SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito: 27/02/2009

Assinatura:_____

Simplificação de Manuais Técnicos de Produtos Tecnológicos para Facilitação da Leitura e Normalização Textual

Fernando Aurélio Martins Muniz

Orientador*: Profa. Dra. Sandra Maria Aluísio*

Monografia apresentada ao Instituto de Ciências Matemáticas e de Computação - ICMC-USP, para o Exame de Qualificação, como parte dos requisitos para obtenção do título de Mestre em Ciências - Ciências de Computação e Matemática Computacional.

USP – São Carlos
Fevereiro de 2009

Simplificação de Manuais Técnicos de Produtos Tecnológicos para Facilitação da Leitura e Normalização Textual

Fernando Aurélio Martins Muniz

Resumo

No Brasil, cerca de 65% da população possui o nível de alfabestismo entre o rudimentar (25%) e o básico (40%). A principal característica que distingue os níveis de alfabetismo é a complexidade do texto. O objetivo da Simplificação Textual é proporcionar ao leitor um melhor entendimento do texto através de sua simplificação. Em tarefas que exigem o uso de documentação técnica, a qualidade da documentação é um ponto crítico, pois caso a documentação seja imprecisa, incompleta ou muito complexa, o custo da tarefa ou até mesmo o risco de acidentes aumenta muito. Manuais de instrução possuem duas relações procedimentais básicas: relação gera, quando uma ação "A" gera automaticamente uma ação "B", e a relação habilita, quando a realização de uma ação "A" permite a realização da ação "B". O projeto aqui descrito está inserido nesse contexto e pretende estudar como as relações procedimentais entre ações *generation* (gera) e *enablement* (habilita) são realizadas em manuais de instruções, dando base para a adaptação do Manual de Simplificação Sintática e do Editor de Simplificação, propostos no projeto PorSimples, para atender ao gênero de textos instrumental. Além disso, pretende-se adaptar as ferramentas de extração de termos para este gênero de texto.

Sumário

Lista de Figuras vi

Lista de Tabelas vii

1 Introdução 1

1.1 Contexto 1

1.2 Motivação e Justificativa 4

1.3 Objetivos do Projeto de Mestrado 6

1.4 Organização da Monografia 6

2 Trabalhos Relacionados: a Linguagem Controlada Usada em Manuais Técnicos, os Sistemas de Geração Automática de Manuais e os Sistemas de Perguntas e Respostas 7

2.1 Linguagem Controlada 7

2.1.1 Definições 7

2.1.2 Os Componentes de uma Linguagem Controlada 9

2.1.3 Plain English 10

2.1.4 Inglês Simplificado AECMA 11

2.1.5 Linguagem Simplificada em Textos do Português do Brasil 12

2.1.6 Sublíngua 13

2.2 Tratamento de Textos Instrucionais em Trabalhos de PLN 14

2.2.1 O Trabalho de Di Eugenio (1992) 14

2.2.2 Geração Multilíngue de Manuais de Instrução 18

2.2.3 Estrutura de Textos Instrucionais e Sistemas de Perguntas & Respostas 27

3 Simplificação Sintática no PorSimples 32

4 Métodos de Extração Automática de Termos 41

4.1 Métodos Linguísticos 41

4.2 Métodos Estatísticos 42

4.3 Métodos Híbridos 44

4.4 O Trabalho de Teline (2004) 44

4.5 O Trabalho de Ribeiro Jr (2008) 46

5 Projeto NorMan 48

5.1 Metodologia Baseada em Corpus para a Extração de Termos de Textos Instrucionais e para a Normalização (Simplificação) de Instruções 48

5.1.1 Compilação e Análise de um Corpus de Manuais Técnicos 48

5.1.2 Proposta de um Método de Extração de Termos Técnicos Sensível ao Gênero Instrucional/Procedimental 49

5.1.3 Adaptação do Manual de Simplificação Sintática e do SIMPLIFICA para Tratarem de Textos Instrucionais/Procedimentais 52

5.2 Tarefas Realizadas e Tarefas Futuras 52

5.3 Cronograma 53

Referências 54

Apêndice A 59

Lista de Figuras

Figura 1 - Os sistemas do autoria e facilitação do projeto PorSimples *3*
Figura 2 - Diagrama de Blocos da Arquitetura de DRAFTER (Paris et al., 1995) *21*
Figura 3 - Exemplos de posturas que um manual de produtos pode adotar (Paris & Scott, 1994) *25*
Figura 4 - Estrutura Global de Manuais de Produtos (Paris & Scott, 1994) *26*
Figura 5 - Precisão obtida pelas três abordagens e a semi-automática (Teline, 2004) *45*
Figura 6 – Cobertura obtida pelas três abordagens e a semi-automática (Teline, 2004) *45*
Figura 7 – Medida F obtida pelas 3 abordagens e a semi-automática (Teline, 2004) *46*
Figura 8 - Expressões linguísticas que serão utilizadas para a extração de termos *50*
Figura 9 - Padrões morfossintáticos para extração de termos compostos *51*
Figura 10 - Padrões morfossintáticos usados na extração de termos simples *51*

Lista de Tabelas

Tabela 1 – Frequência das formas gramaticais em português para a relação gera (Delin et al., 1994).....27
Tabela 2 - Frequência das formas gramaticais em português para a relação habilita (Delin et al., 1994).....27
Tabela 3 - Cronograma de atividades.....53

1 Introdução

1.1 Contexto

Esse projeto está sendo desenvolvido no *Núcleo Interinstitucional de Lingüística Computacional* (NILC[1]), criado em 1993, sendo ele um grupo interdisciplinar dedicado à pesquisa e ao desenvolvimento de recursos, ferramentas e sistemas de Processamento de Língua Natural (PLN), como o desenvolvimento de corpus e léxicos, etiquetadores morfossintáticos, sumarizadores automáticos, aplicações de tradução de máquina e ferramentas de suporte à escrita, entre outros. Alguns projetos do NILC são de fundamental importância para este trabalho, como o projeto PorSimples[2] e o ExPorTer[3], descritos abaixo.

O termo letramento vem sendo usado no Brasil para designar o conceito de alfabetismo funcional, que é a capacidade das pessoas de utilizar efetivamente a leitura e a escrita diante de vários aspectos da sua vida social (Ribeiro, 2006). Desde 2001, o índice INAF [4] (Indicador de Alfabetismo Funcional) tem medido os níveis de analfabetismo funcional na população Brasileira, e ele identifica quatro níveis de letramento na população (dados de 2007):

> 1 – Analfabetismo: Corresponde a situação daqueles que não conseguem realizar tarefas simples que envolvem a leitura de palavras e frases. Esse nível abrange 7% da população Brasileira;
>
> 2 – Alfabetismo nível rudimentar: Corresponde à capacidade de localizar uma informação explícita em textos curtos e familiares, como anúncios ou pequenas cartas. Este nível atinge 25% da população Brasileira;
>
> 3 – Alfabetismo nível básico: Neste nível, a pessoa tem a habilidade de ler e compreender textos de tamanho médio e podem localizar informações e realizar inferências simples. Este nível abrange 40% da população Brasileira;

[1] http://nilc.icmc.usp.br/
[2] http://caravelas.icmc.usp.br/wiki/index.php/Principal
[3] http://nilc.icmc.usp.br/nilc/projects/termextract.htm
[4] http://www.ipm.org.br/

4 – Alfabetismo nível pleno: Corresponde a habilidade de ler textos longos, encontrar vários tipos de informação, comparar textos diferentes e realizar inferências. Este nível corresponde a 28% da população Brasileira.

A principal característica distintiva entre os três níveis de letramento é complexidade do texto, que nos remete para área de Simplificação Textual. Simplificação textual é uma área de pesquisa emergente no campo de Processamento de Linguagem Natural (PLN). O objetivo da Simplificação Textual é proporcionar ao leitor um melhor entendimento do texto através de sua simplificação.

Algumas abordagens envolvem simplificação da estrutura léxica e sintática, através da substituição de palavras incomuns por palavras mais comuns e freqüentes (simplificação lexical – SL) e também pela divisão e mudança da sintaxe da sentença (Mapleson, 2006) (Max, 2006).

O exemplo a seguir faz uma simplificação sintática de uma oração em voz passiva transformando a sentença em voz ativa:

Original: *Essa escolha deve ser feita eventualmente por Lula.*

Simplificado: *Lula deve, eventualmente, fazer essa escolha*

Outro próximo exemplo faz a simplificação sintática fazendo a divisão de uma oração coordenada sindéticas em várias sentenças, através da eliminação do marcador de coordenação e repetindo o núcleo do sujeito para cada período:

Original: *A Mesa tem de aprovar o parecer e enviá-lo ao plenário.*

Simplificado: *A Mesa tem de aprovar o parecer. A Mesa tem de enviá-lo ao plenário.*

Outras abordagens envolvem o descarte de partes do texto (Petersen *et al*., 2007) e a adição de material extra para explicar termos complexos. Esta última abordagem é conhecida como elaboração lexical (EL) que consiste no acréscimo de definições, sinônimos, antônimos ou hiperônimos destas palavras, ao invés de substituí-las (Urano, 1998). Por exemplo, para a sentença "Todos sabem que Ken é diligente e gentil com os outros." uma SL a transformaria em "Todos sabem que Ken é esforçado e gentil com os outros." enquanto que um exemplo de EL seria: "Todos sabem que Ken é diligente, ou esforçado, e gentil com os outros."

O projeto PorSimples (Simplificação Textual do Português para Inclusão e Acessibilidade Digital) aborda a simplificação de textos visando a construção de sistemas que

promovem o acesso a textos em Português do Brasil para pessoas com nível rudimentar e básico de letramento e também para aquelas com problemas cognitivos, como dislexia e afasia. Dois sistemas estão sendo desenvolvidos: O primeiro é um sistema de **autoria on-line** para ajudar na produção de textos simples e o segundo é um sistema de **facilitação de textos** que já estão prontos, incluindo as tarefas de sumarização, simplificação e explicitação visual de características do texto como as entidades nomeadas, os argumentos dos verbos que ajudam a responder as perguntas quem, como, o que, onde, entre outras e também as relações entre partes do texto que tem base nos marcadores discursivos de contraste, adição, exemplificação, entre outros. A interação entre os sistemas pode ser observada na Figura 1.

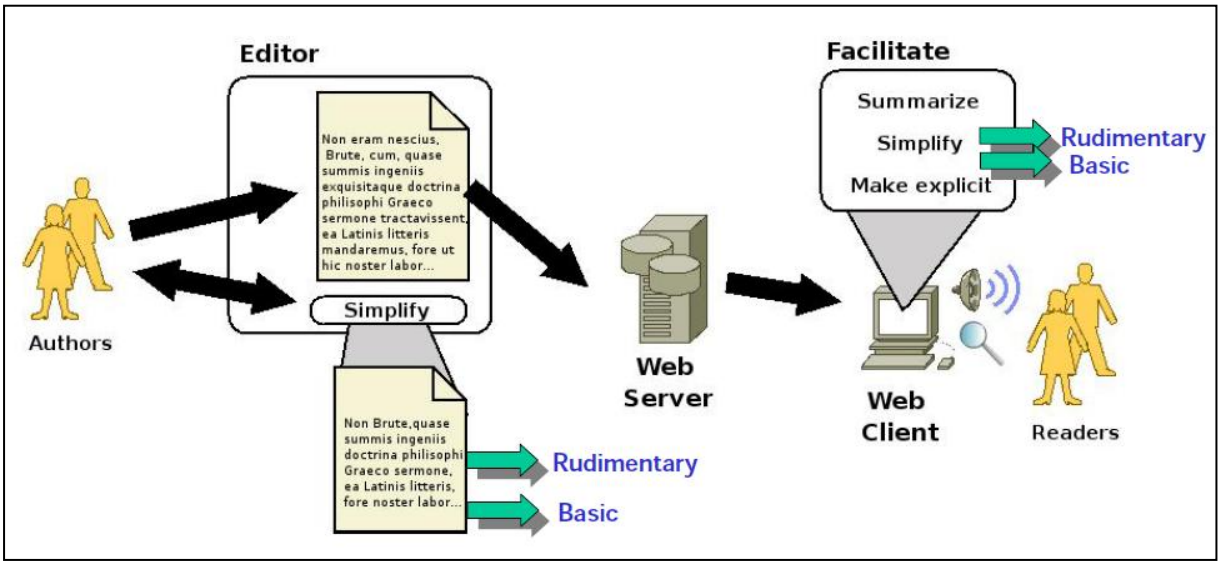

Figura 1 - Os sistemas do autoria e facilitação do projeto PorSimples

O foco principal do projeto PorSimples são os textos do gênero informativo, diferente deste projeto, que visa o estudo e a adaptação do sistema de autoria on-line para a escrita de textos do gênero instrucional, particularmente manuais técnicos de produtos tecnológicos.

Textos instrucionais, também chamados de textos procedimentais, consistem de uma seqüência de instruções, concebidas com certa precisão a fim de atingir um objetivo (por exemplo, montar um computador). Textos procedimentais explicam como realizar um determinado objetivo através de ações que estão ao menos parcialmente temporalmente organizadas. Textos instrucionais consistem de seqüências de instruções concebidas de forma a alcançar um objetivo. O usuário deve seguir passo a passo as instruções a fim de alcançar os resultados esperados (Aouladomar & Saint-Dizier, 2005).

No projeto PorSimples, foi compilada uma lista de palavras simples contendo palavras consideradas simples para os jovens (Biderman, 2005) e também uma lista de palavras concretas (Janczura *et al*., 2007).

Essas palavras auxiliam os sistemas do PorSimples a identificarem os possíveis casos de simplificação léxica presentes nos textos. No caso dos manuais de instruções, existem termos que devem ser preservados, e para isso será necessário a criação de uma ferramenta para extração automática de termos em manuais. Este tipo de extração é novidade, pois no melhor do nosso conhecimento, não há ferramentas de extração dedicada a textos de manuais.

Existem três tipos de abordagens para extração de termos. A primeira, chamada de lingüística, faz uso apenas de conhecimento lingüístico. A segunda utiliza apenas métodos estatísticos, e é denominada abordagem estatística. A última abordagem, que é a mais utilizada, combina conhecimentos lingüísticos e estatísticos e é denominada abordagem híbrida (Teline, 2004). Essas abordagens foram apresentadas no ExPorTer serão explicadas com mais detalhes no Capítulo 4, que também apresentará a abordagem desenvolvida por (Ribeiro Jr, 2008), que utiliza conhecimento semântico para auxiliar a tarefa de extração de termos.

Além disso, no projeto PorSimples foi produzido um manual de simplificação sintática, que será apresentado no Capítulo 3 deste trabalho, e que deve ser adaptado para tratar mais especificamente das formas lingüísticas utilizadas no português para expressar de forma efetiva as relações **gera** e **habilita**, largamente usadas nos manuais e explicadas na Seção 1.2

1.2 Motivação e Justificativa

Tarefas operacionais, procedimentos de manutenção e diagnósticos de falhas em sistemas técnicos complexos requerem o uso de documentação técnica. A qualidade dessa documentação é um ponto crítico. Se a documentação está imprecisa, incompleta ou difícil de entender, o custo e o tempo da operação de reparo irá aumentar muito. Até mesmo prejuízo a equipamentos caros ou acidentes com vítimas humanas podem ocorrer devido ao mau entendimento da documentação técnica (Eijk, 1997).

O desenvolvimento contínuo de novas tecnologias e produtos combinados com o fato de que grande parte da população tem um nível básico e rudimentar de letramento torna clara

a importância da boa escrita de manuais técnicos bem como demonstra ser um assunto interessante a ser estudado.

Textos procedimentais/instrucionais consistem de uma seqüência de instruções concebidas com certa precisão a fim de alcançar um objetivo. O leitor deve seguir passo a passo cuidadosamente as instruções fornecidas pelo manual a fim de alcançar o objetivo (Fontan & Saint-Dizier, 2008).

Em (Paris *et al.*, 1995) uma análise de requisitos para uma ferramenta de suporte à escrita de documentos técnicos multilíngüe confirmou que uma ferramenta de auxílio a escrita é mais útil do que uma ferramenta de geração automática que mantém o escritor longe do texto produzido.

(Paris *et al.*, 1994) mostra que os manuais de instruções podem ter diferentes estilos, nem todas as instruções são registradas usam uma seqüência de imperativos, como seria mais natural de se esperar, e que diferentes partes do manual freqüentemente usam diferentes estilos.

(Aouladomar, 2005a) faz uma análise da estrutura de manuais e de perguntas relacionadas a textos procedimentais (por exemplo: "Como?" e "Por quê?") e mostra que perguntas e fragmentos de textos procedimentais podem ser combinados a fim de produzirem respostas para máquinas de busca.

O Filósofo Alvin Goldman identificou duas relações procedimentais básicas, *gera* e *habilita* (Goldman, 1970 apud Delin, 1994). A relação gera é uma relação que aparece entre duas ações e que passa o sentido de que a após a realização da ação "A", a ação "B" ocorrerá automaticamente, ou seja, "A" gera "B". No português, expressões lingüísticas da relação gera geralmente envolvem o conectivo "para", primeiramente seguido por um infinitivo e, em ocasiões raras, seguido por um sintagma nominal. O seguinte trecho de um manual de instruções, em português, de uma serra elétrica exemplifica essa relação (Delin, 1994)

> *Para colocar a serra na posição de corte oblíquo, solte a porca borboleta e incline a sapata para o ângulo desejado.* (Black&Decker)

A relação habilita é o tipo de relação que ocorre quando a realização de uma ação "A" não resulta na realização automática da ação "B". Apesar do conectivo "para" também ser usado para a relação habilita, ele não foi encontrado em (Delin, 1994). Ao contrário, neste

estudo, as relações habilita foram encontradas através de sinais de ordem temporal nas ações envolvidas, em orações consecutivas ou ligadas pela conjunção “e”.

Desligue a serra da tomada antes de fazer qualquer ajuste. (Black&Decker)

1.3 Objetivos do Projeto de Mestrado

O objetivo deste trabalho é estudar, via um corpus criado no escopo desta pesquisa, como as relações procedimentais entre ações *generation* (gera) e *enablement* (habilita) são realizadas em instruções de manuais técnicos de produtos tecnológicos em português. Este estudo dará base para uma adaptação no Manual de Simplificação Sintática e no Editor de Simplificação, propostos no projeto PorSimples para atender ao gênero de textos instrumental.

Este projeto propõe também a adaptação das ferramentas de extração de termos para este gênero de textos, via uma extração de termos dedicada às expressões lingüísticas das relações gera e habilita.

1.4 Organização da Monografia

Esta monografia está organizada conforme o que se segue. No Capítulo 2 são apresentados trabalhos relacionados com a criação e uso de Linguagens Controladas, que são amplamente utilizadas na produção de manuais técnicos. Também, são apresentados trabalhos da área de PLN sobre modelagem computacional de instruções, sistemas de geração automática de manuais e sistemas de perguntas e respostas adaptados ao tratamento de textos instrucionais. No Capítulo 3 é apresentado o Manual de Simplificação Sintática do projeto PorSimples e sua aplicação em uma ferramenta de simplificação simbólica desenvolvida no projeto. No Capítulo 4 é apresentada uma comparação sobre as várias abordagens dos métodos de extração automática de termos. No Capítulo 5 é apresentada a proposta deste projeto juntamente com o cronograma de desenvolvimento das várias tarefas para sua execução.

2 Trabalhos Relacionados: a Linguagem Controlada Usada em Manuais Técnicos, os Sistemas de Geração Automática de Manuais e os Sistemas de Perguntas e Respostas

2.1 Linguagem Controlada

Os princípios da linguagem controlada foram originados pelo inglês básico de Ogden (1932), na década de 30. O Inglês básico foi proposto tanto como uma língua internacional quanto uma base para o aprendizado do inglês como segunda língua. Ela consiste de 850 palavras básicas (Veja Apêndice A) e algumas inflexões e regras de derivação (Kaji, 1999).

A primeira linguagem controlada para documentação técnica, chamada *Caterpillar Fundamental English* (CFE), foi desenvolvida na década de 60 para melhorar a compreensão de leitura dos usuários não nativos dos manuais da Caterpillar, que é uma fabricante de equipamento pesado. Além disso, linguagens controladas têm mantido uma forte associação com o objetivo de reduzir custos e tempos de tradução de manuais. A Caterpillar, por exemplo, desenvolveu nos anos 90 uma linguagem mais controlada ainda que a CFE (chamada *Caterpillar Technical English* (CTE)) em conjunto com o Center for Machine translation (CMT) da Carnegie Mellon University[5] para facilitar a tradução automática (Kaji, 1999).

Várias outras companhias internacionais como a Ericsson, IBM, Xerox e Scania seguiram os passos da Caterpillar, gerando várias linguagens controladas ou simplificadas.

Como resultado dos crescentes requisitos de qualidade e da globalização da economia, o uso de linguagens controladas tem sido aumentado nas documentações técnicas, e junto com isso cresce o interesse do uso de PLN para apoiar a criação de textos fontes seguindo as recomendações de uma linguagem controlada (Eijk, 1997).

2.1.1 Definições

Usada em documentação técnica, linguagem controlada é uma linguagem escrita de acordo com regras rigorosas e com um vocabulário limitado. O objetivo é eliminar a possibilidade de ambigüidade. Foi verificado que textos fontes escritos numa linguagem controlada são muito mais fáceis de serem processados em tradução de máquina (Edwards & Kingscott, 1997).

[5] Veja detalhes em: Projeto KANT - http://www.lti.cs.cmu.edu/Research/Kant

Os fatores importantes em uma análise do processo de compreensão de um texto são, segundo Leffa (1996): o texto, o leitor e as circunstâncias em que se dá o encontro. Entre os fatores relativos ao texto, destacam-se, a **legibilidade** (apresentação gráfica do texto) e a **inteligibilidade** (uso de palavras freqüentes e estruturas sintáticas menos complexas). As linguagens controladas surgiram para amenizar o problema da inteligibilidade de textos, ao reduzirem a complexidade da estrutura sintática do texto, ao realizarem a desambiguação lexical e ao tratarem de problemas de tradução, pois o controle semântico e sintático facilita a tradução de uma língua para outra. Podemos citar as seguintes vantagens no uso de linguagens controladas (Altwarg, 2006; Burg, 1996; Cole, 1997;):

- Melhoria na inteligibilidade, pela redução da ambigüidade lexical e estrutural, e por prescrever regras estilísticas
- Melhoria na manutenibilidade, que é conseqüência direta da melhoria na inteligibilidade
- Processamento computacional facilitado, em função da ambigüidade reduzida e das regras estilísticas prescritas
- Tradução automática facilitada

Existem também algumas desvantagens que devem ser conhecidas, segundo (Altwarg, 2006; Burg, 1996; Cole, 1997;):

- Redução do poder de expressão, causada por restrições da linguagem
- Redução na velocidade de escrita, pois é necessário "pensar" mais para decidir que palavras e construções sintáticas são permitidas para serem usadas. A correção das violações das regras da linguagem também custa tempo.
- Considerável investimento, necessário para o treinamento das pessoas envolvidas para escreverem na linguagem controlada corretamente.

As linguagens controladas são mais usadas em aplicações comerciais e industriais, como na autoria de manuais técnicos de uso e manutenção, que possuem uma grande quantidade de documentos complexos que são freqüentemente atualizados nos quais a terminologia é específica de um domínio. O melhor exemplo conhecido de linguagem controlada é o *AECMA Simplified English*[6], uma norma internacional aceita para escrever manuais técnicos na indústria aeroespacial. O padrão AECMA teve início em 1979, e cresceu

[6] AECMA é o acrônimo francês para Associação Européia de Fabricantes de Equipamentos Aeroespaciais

da experiência coletiva de várias grandes fabricantes nas últimas décadas, cujo objetivo é simplificar a documentação técnica, tanto para leitura na língua original, quanto para facilitar a tradução automática e para exportação para outros mercados (Mitkov, 2003). Além disso, encontramos o uso de linguagem controlada em domínios nos quais a documentação é tradicionalmente muito complexa ou mal escrita, como nos documentos do governo, do setor de finanças e do jurídico.

2.1.2 Os Componentes de uma Linguagem Controlada

Uma linguagem controlada pode ser definida como uma série de restrições de vocabulário, gramática e estilo. Uma linguagem controlada prescreve esses elementos de uma maneira limitada e formal. A restrição de vocabulário é de fundamental importância para qualquer linguagem controlada, tendo como principal objetivo a redução da ambigüidade (Kaji, 1999). Além da restrição de vocabulário, um estudo terminológico deve ser realizado dentro do domínio específico. Através desse estudo, serão identificados candidatos a termos e suas variantes, que serão usados na criação de uma base de termos. Essa base é utilizada para evitar que informações importantes sejam perdidas durante a simplificação de uma frase. Toda linguagem controlada possui regras gramaticais específicas. Essas regras são responsáveis por definir quais tipos de construções de frase são permitidas. Geralmente essas regras visam melhorar os textos propondo frases simples e curtas. Alguns exemplos de regras gramaticais podem ser observados a seguir (Kaji, 1999):

- Usar frases curtas. Esta regra é simples, porém muito eficaz, pois frases curtas contêm menos ambigüidades.
- Eliminar palavras redundantes. Também reduz a ambigüidade e ajuda na tradução automática.
- Evitar o uso de voz passiva.
- Não usar seqüências muito grandes de substantivo.

As gramáticas devem ser adaptadas para cada linguagem controlada conforme suas necessidades. Formato e layout também fazem parte da linguagem controlada. Por exemplo, a utilização de uma lista pode eliminar uma estrutura coordenada complexa, e resulta em pequenas frases, muito mais fáceis de serem interpretadas. A padronização de convenções estilísticas também deve ser aplicada no formato de datas e moedas, tabelas, e variantes ortográficas (Kaji, 1999).

2.1.3 Plain English

O *Plain English* é um movimento na Inglaterra e Estados Unidos que se iniciou no fim dos anos 70 como uma reação à falta de clareza na linguagem usada nos documentos do governo e comércio. Ele usa recomendações (a *Plain Language*[7]) que, em princípio pode ser aplicado a qualquer língua natural.

A *Plain Language* pode ajudar a escrever um texto simples e curto, claro, efetivo e fácil de entender para o público alvo. Um documento que siga a *Plain Language* usa palavras economicamente e em um nível que a audiência possa entender, criando um texto sensível ao leitor como reforça Leffa (1996):, acima.

Algumas recomendações de como escrever e organizar as informações em *Plain Language*[8] são:

a) Escreva utilizando pronomes pessoais para falar diretamente aos leitores;

b) Seja direto; elimine qualquer ambigüidade;

c) Use a lógica simples e comum criando conexões entre idéias obvias;

d) Use títulos ou subtítulos que informam ou resumem o texto;

e) Retire toda a informação que não é essencial para o propósito do texto, evitando assim longas descrições nominais;

f) Priorize a informação e coloque o que for mais importante no início;

g) Use gráficos, planilhas, e imagens para reforçar pontos e fatos complicados;

h) Use índices/sumário para documentos grandes ou crie uma introdução curta com o conteúdo de cada item.

Outras recomendações para reescrever e revisar textos são:

i) Mantenha o sujeito e o verbo próximos do início da sentença;
ii) Explique somente uma idéia por sentença;
iii) Faça frases entre 25 a 35 palavras em média (dados para o inglês);
iv) Use verbos em vez de substantivos para exemplificar sua ação;

[7] http://www.plainlanguage.gov/
[8] http://www.plainlanguagenetwork.org/

v) Use a voz ativa: tendo certeza que o autor é bem definido como sua ação;

vi) Use palavras e construa sentenças positivas; evite frases com aspectos negativos;

vii) Faça uma sintaxe (construção gramatical) simples;

viii) Evite a formalidade quando desnecessário; Simplifique suas palavras, escolha a linguagem cotidiana;

ix) Evite jargões, palavras estrangeiras e termos jurídicos;

x) Use palavras técnicas com cuidado: defina ou providencie exemplos descritivos;

xi) Minimize abreviações;

xii) Coloque a idéia principal antes das exceções e condições;

xiii) Use termos e conceitos importantes de maneira consistente.

Entretanto, mesmo as recomendações da *Plain Language* não especificam detalhadamente o que seria uma sintaxe simples nem o que seria a linguagem cotidiana, embora o manual com as recomendações apresente vários exemplos de simplificação para a língua inglesa.

2.1.4 Inglês Simplificado AECMA

O inglês simplificado AECMA é utilizado pela maioria das grandes fabricantes de equipamento aeroespacial, e por muitas empresas aéreas. O guia de inglês simplificado especifica três fontes de palavras:

1. Cerca de 950 palavras básicas aprovadas, todas com definições não técnicas bem definidas; entre elas incluem todas as preposições importantes, artigos e conjunções, bem como substantivos básicos, verbos, adjetivos e advérbios.
2. Um número ilimitado de nomes técnicos, divididos em vinte categorias, que podem ser escolhidas pela organização, mas usadas somente como adjetivos ou substantivos, de acordo com certas orientações.
3. Verbos técnicos para designar seis categorias de processos de fabricação especificados pelo usuário, sujeito a regras rígidas de utilização.

Este padrão de inglês simplificado tem cerca de cinqüenta e cinco regras que regem o uso de palavras e a construção de frases. Algumas delas são muito precisas:

> *Você deve quebrar grupos de substantivos de quatros ou mais palavras reescrevendo-os, usando hifenização ou uma combinação dos dois.*

Entre as regras precisas, existem várias sobre pontuação. Algumas outras são um pouco vagas, como:

Mantenha um tópico por frase

ou então expressam metas de escritas desejáveis:

Tente variar o tamanho de frases e construções para manter o texto interessante.

A maioria das regras vagas ou orientações de metas podem ser vistas como princípios que se aplicam à boa prática de escrita de forma geral (Mitkov, 2003).

2.1.5 Linguagem Simplificada em Textos do Português do Brasil

Há uma tendência atual pelo governo brasileiro da criação de guias ou cartilhas[9] que utilizam uma linguagem simples e acessível a pessoas de qualquer idade e com um menor grau de instrução. Também, a RESOLUÇÃO No 7, DE 29 DE JULHO DE 2002[10] estabelece regras e diretrizes para os sites da Administração Pública Federal que ditam que as páginas devem ser de fácil legibilidade; apresentar os conteúdos com clareza, simplicidade, objetividade, organicidade, atualidade, e veracidade; usar linguagem simples e direta, especialmente nas páginas iniciais; além de outros critérios.

No meio jurídico, a campanha Simplificação da Linguagem Jurídica, lançada pela Associação dos Magistrados Brasileiros (AMB) em 2005 foi criada para que magistrados, advogados, promotores e outros operadores do Direito utilizem linguagem mais simples para facilitar o entendimento da população em geral. Nesta campanha foi lançado um livro contendo um glossário com as expressões técnicas mais usuais nos meios forenses e as respectivas traduções, na tentativa de combater o juridiquês. Geralmente estas cartilhas de orientação legal (COL) tornam as leis mais acessíveis ao desconstruir a linguagem jurídica. A versão em forma de cartilha ilustrada do Novo Código Civil Brasileiro que entrou em vigor em 11/1/2003 foi lançada em 2003 (Carvalho Netto, 2003).

[9] Veja, por exemplo, a cartilha Lei Maria da Penha em: http://www.agenciaalagoas.al.gov.br/noticias-governo/governo-lanca-cartilha-sobre-a-lei-maria-da-penha (Acessado em 6/2/2008) ou a cartilha lançada pelo Ministério do Trabalho e Emprego com informações úteis para brasileiros e brasileiras que vivem no exterior: http://download.uol.com.br/ultnot/cartilha.pdf.

[10] http://www.icpbrasil.gov.br/e-gov/Resolucao_N_7_Diretrizes_para_Sitios_Internet_29_07_02.pdf

Outras cartilhas lançadas pelo governo com informações de interesse amplo como saúde, transito, etc. também são escritas em linguagem simples. Uma em específico foi lançada pelo Ministério do Trabalho e Emprego com informações úteis para brasileiros que vivem no exterior[11].

O governo brasileiro também editou leis que tornam as bulas de medicamentos mais fáceis de serem lidas, separando as informações para o público geral da informação para médicos e farmacêuticos e lançando o site de busca Bulário da Anvisa[12].

Na mídia jornalística, há um interesse grande pela linguagem simplificada, aproveitando jornalistas experientes e bem articulados para a criação de seções como "De palavra em palavra", do jornal Estado de São Paulo "Para seu Filho Ler" do Jornal Zero Hora entre outras.

A ciência também foi desmistificada para crianças com revistas como "Ciência Hoje das Crianças"[13] do **Instituto Ciência Hoje** (ICH), uma organização social de interesse público sem fins lucrativos vinculada à Sociedade Brasileira para o Progresso da Ciência (SBPC). A revista **Ciência Hoje das Crianças** mostra ao público infantil que a ciência faz parte da vida de cada um e pode ser muito divertida. A revista estimula a curiosidade e a compreensão dos fenômenos do dia-a-dia, com a ajuda de ilustrações e experiências que podem ser realizadas pelas próprias crianças. É instrumento fundamental em sala de aula como fonte de pesquisa aos professores e de grande importância para os alunos na elaboração de deveres e projetos escolares. A publicação é adotada pelo MEC e distribuída para 107 mil escolas, como material de apoio paradidático.

Uma das vantagens dos materiais simplificados é que preparam o leitor para a leitura eventual dos textos autênticos. Por isso, segundo Ramos (2006) os textos instrucionais sempre devem dispor da referência bibliográfica ao final das unidades para que os leitores as consultem quando for de seu interesse e necessidade.

2.1.6 Sublíngua

Uma sublíngua é uma língua usada em domínios ou assuntos específicos, tais como biologia e engenharia aeroespacial. Exemplos de sublínguas são a linguagem de relatórios

[11] http://download.uol.com.br/ultnot/cartilha.pdf
[12] http://bulario.bvs.br/
[13] http://ich.unito.com.br/

meteorológicos, manuais de manutenção de aviões, artigos científicos sobre fármacos e anúncios de venda de imóveis (Grishman, 2001).

Por exemplo, a palavra ‘chip’ no domínio da computação significa um pedaço de material semicondutor, enquanto ‘chip’ em um bar britânico significa um pedaço de batata frita. Cada um destes domínios tem um vocabulário para suas necessidades. A principal diferença entre uma linguagem controlada e uma sublíngua é que na linguagem controlada os termos, a sintaxe e a semântica são propositalmente limitados, geralmente com objetivos particulares em mente, e na sublíngua essas limitações não são especificadas e evoluem naturalmente (Altwarg, 2006)

2.2 Tratamento de Textos Instrucionais em Trabalhos de PLN

Nesta seção, apresentamos trabalhos de análise de corpus de manuais de instrução em várias línguas para: a criação de modelos computacionais de instrução (Seção 2.2.1); a criação de uma ferramenta, baseada na tecnologia de geração multilíngüe, que dá suporte à produção de manuais de usuário de produtos de software (Seção 2.2.2); a criação de sistemas de perguntas e respostas baseados na estrutura global de textos instrucionais/procedimentais (Seção 2.2.3).

Tanto o trabalho de Di Eugenio (1992), apresentado na Seção 2.2.1, como o do grupo de pesquisadores que desenvolveram o DRAFTER (Delin *et al.*, 1994), apresentado na Seção 2.2.2, realizaram um estudo das relações procedimentais gera (*generation*) e habilita (*enablement*). Estas relações são importantes neste projeto de mestrado, pois elas são fundamentais para a escrita adequada de manuais e sua realização varia de língua para língua, indicando realizações preferenciais que, ao serem levadas em consideração na escrita de manuais alcançarão uma eficácia maior para o uso, manutenção e cuidados dos produtos.

2.2.1 O Trabalho de Di Eugenio (1992)

No trabalho de Barbara Di Eugenio (1992) é feita uma análise da **cláusula de propósito ou finalidade** no contexto do entendimento de instruções. Essa análise mostra que os objetivos afetam a interpretação ou execução das ações, e dão apoio ao uso do modelo de relações **gera** (*generation*) e **habilita** (*enablement*) entre as ações.

A análise das cláusulas de finalidade é relevante para o problema de entendimento de instruções em língua natural por que elas:

- estabelecem explicitamente objetivos e sua interpretação mostra que o objetivo que o ouvinte adota o guia durante a realização das tarefas.
- aparecem para expressar as relações entre ações gera ou habilita, apoiando a proposta de que essas duas relações são necessárias para o modelo de ações.

Segundo Di Eugenio (1992), uma pessoa A usa uma cláusula de finalidade para explicar um objetivo X a uma pessoa B através da execução de uma tarefa Y. Porém, um ponto importante que tem sido ignorado é que o objetivo X também limita a interpretação da tarefa Y. Por exemplo, considere o exemplo abaixo, tomado de Di Eugenio (1992), com traduções para o português entre parênteses:

> ***a) Place a plank between two ladders (Coloque uma prancha entre duas escadas)***
>
> ***b) Place a plank between two ladders to create a simple scaffold (Coloque uma prancha entre duas escadas para criar um simples andaime)***

Nos dois casos (a) e (b) a ação a ser executada é "Coloque *uma prancha entre duas escadas*". Porém, essa informação não é especifica, e existem várias maneiras de "*colocar a prancha*". O objetivo "*criar um simples andaime*" restringe a escolha da posição de colocação da prancha. A sentença no infinitivo "*to create a simple scaffold*"/"*para criar um simples andaime*" é uma **cláusula de propósito** que expressa o propósito de um agente na realização de uma dada ação.

Na maioria dos casos, um objetivo X descreve uma mudança no mundo. No entanto, em alguns casos:

1) a mudança não é no mundo, mas sim no conhecimento de quem recebe a instrução. Ao executar uma tarefa Y, o conhecimento de quem a está executando pode ser mudado em respeito a certa proposição ou ao valor de uma certa entidade. Por exemplo:

> *You may want to hang a coordinating border around the room at the top of the walls.* ***To determine the amount of border****, measure the width (infeet) of all walls to be covered and divide by three. Since borders are sold by the yard, this will give you the number of yards needed.*
>
> *(Você pode querer pendurar uma moldura no topo da parede ao redor da sala.* ***Para determinar a quantidade de moldura****, meça a largura de todas as paredes a*

serem cobertas e divida por três. Como as molduras são vendidas em jardas, isto lhe dará o número de jardas necessárias.)

Vários desses exemplos envolvem verbos como "*check*", "*make sure*" (checar, ter certeza, entre outros) seguido pelo complemento "**that" (que)** descrevendo descreve um estado **e**. O uso de tais verbos tem um efeito pragmático que faz com que o executor da tarefa não somente cheque se **e** acontece, pois se **e** não acontecer, ele fará algo para funcionar. Por exemplo:

To attach the wires to the new switch, use the paper clip to move the spring type clip aside and slip the wire into place. Tug gently on each wire ***to make sure it's secure****.*

(Para fixar os fios no novo interruptor, use um clip de papel para mover o grampo de mola para o lado e colocar o fio no lugar. Puxe suavemente cada fio ***para ter certeza que está seguro.****)*

2) A cláusula de finalidade pode informar ao executor que o mundo **não** deve mudar, que dado evento deve ser evitado. Por exemplo:

Tape raw edges of fabric to prevent threads from raveling as you work.

(Costure as bordas de tecido cru para evitar que o fio desfie enquanto você trabalha.)

Do ponto de vista do processamento do discurso, interpretar uma cláusula de finalidade pode afetar o modelo de discurso, em particular por introduzir novos referentes. Isto acontece quando o efeito de Y é criar um novo objeto, e X o identifica. Verbos freqüentemente usados neste contexto são ***criar, fazer, formar***, etc. Por exemplo:

Join the short ends of the hat band to form a circle.

(Una as extremidades curtas da banda do chapéu para formar um círculo.)

2.2.1.1 As Relações Gera e Habilita

A relação entre duas ações (Ação Y contribui para o objetivo X), na maioria das vezes expressa uma relação gera, e no resto, habilita.

Gera é a relação entre ações ligada pela preposição "***by***" (por/via) em Inglês – "*turning on the light by flipping the switch*" *(acender a luz via aperto do interruptor).* Formalmente, pode-se dizer que uma ação X condicionalmente gera outra ação Y se e somente se:

- X e Y são simultâneos;
- X não é parte da execução de Y;
- Quando X acontece, uma série de condições C tornam-se válidas, assim a junção da ocorrência de X e C implicam na ocorrência de Y. No caso da relação gera entre flipping the switch e turning on the light, o conjunto C de condições seriam: fiação, interruptor e lâmpada estarem funcionando.

Gera é uma relação pervasiva entre descrições de ações naturais. No corpus do trabalho de Barbara Di Eugenio, a cláusula encabeçada por "*by*" (por) é usada com menos freqüência do que as cláusulas de finalidade para expressar a relação gera: 95% de 101 cláusulas de finalidade expressaram relação gera, contra 27 cláusulas "por". Isto significa que a relação gera em textos instrumentais é expressa principalmente por cláusulas de finalidade, na língua inglesa. Estes mesmos resultados são confirmados pelo trabalho de Delin *et al*. (1994), mostrado na Seção 2.2.2.

A relação de **habilita** acontece entre duas ações X e Y só e somente se uma ocorrência de X traz certas condições que são necessárias (mas não necessariamente suficientes) para o desempenho subseqüente de Y. Apenas cerca de 5% dos exemplos do trabalho de Barbara Di Eugenio demonstram a relação de habilita:

Unscrew the protective plate to expose the box. (Desaperte os parafusos da placa de proteção para expor a caixa.)

Desapertar os parafusos da placa de proteção **habilita** a *remoção da placa que* **gera** a *exposição da caixa*.

As relações **gera** e **habilita** são necessárias para modelar ações. Um das justificativas para isto é que elas nos permitem chegar a conclusões sobre a execução de ações.

Se uma ação **X gera Y**, duas ações foram descritas, mas apenas X, o gerador, precisa ser realizada. Em contraste à relação gera, se **X habilita Y**, após executar X, Y ainda precisa ser executada: X deve temporariamente preceder X, no sentido de que X deve começar, mas

não necessariamente terminar, antes de Y. No exemplo a seguir, ***hold (segurar)*** deve continuar por toda duração de ***fill (encher):***

> *Hold the cup under the spigot to fill it with coffee. (Segure a xícara embaixo da torneira para enchê-la com café.)*

Do mesmo jeito que a relação gera afeta a execução do gerador, a ação habilita afeta a execução da ação habilita.

2.2.2 Geração Multilíngüe de Manuais de Instrução

O Instituto de Pesquisa em Tecnologia da Informação (ITRI[14]) existiu como um departamento da Universidade de Brighton até 2005, entretanto, a pesquisa em Processamento de Linguagem Natural continua nesta universidade com o grupo Natural Language Technology Group (NLTG[15]). Dentre as várias pesquisas desenvolvidas tanto no ITRI como no NLTG, neste projeto nos interessa as que envolvem estudos de corpus de textos instrucionais e os projetos relacionados a este gênero de textos. Dois projetos são importantes nesta pesquisa: DRAFTER (DRafting Assistant For TEchnical wRiters) e AGILE (Automatic Generation of Instructions in Languages of Eastern Europe).

DRAFTER (DRafting Assistant For TEchnical wRiters) (Paris and Scott, 1994; Delin *et al*., 1994; Paris *et al*., 1995) foi um projeto de 4 anos de duração que utilizou a tecnologia da geração multilíngüe para dar suporte na produção de manuais de usuário de produtos de software em francês e em inglês. AGILE utilizou a tecnologia para geração de manuais de software em múltiplas línguas sem a necessidade de tradução, desenvolvida no projeto DRAFTER, para adicionar 3 novas línguas ao sistema de geração: búlgaro, russo e tcheco.

Diferente dos sistemas de geração automática de manuais de instrução, que deixam os autores fora do processo, focando na produção automática das instruções, o DRAFTER é uma ferramenta de apoio destinada a ser integrada ao ambiente de trabalho dos autores, visando o auxílio da produção e a automatização de algumas tarefas mais tediosas.

O primeiro passo para desenvolver o DRAFTER, foi a condução de uma análise de requisitos dos usuários, identificando o ambiente amplo no qual a ferramenta proposta seria usada. Nessa análise de requisito, foram feitas entrevistas com autores técnicos, tratando de

[14] http://www.itri.brighton.ac.uk/aboutindex.html
[15] http://www.nltg.brighton.ac.uk/nltg/

vários tópicos, desde as limitações de tempo e orçamento para essa área de trabalho, como também as partes do trabalho que eram consideradas interessantes, a dificuldade, entre outros. Os entrevistados explicaram a sucessão e a divisão do tempo na tarefa de criação de documentação. Além disso, a forma como os documentos eram desenvolvidos, as fontes de informações, e as tarefas e recursos utilizados foram identificados.

Os autores do sistema relatam que uma das descobertas durante a análise de requisitos foi a de que os autores de manuais gastam pouco tempo escrevendo textos novos. Grande parte do trabalho deles é atualizar os documentos existentes. Além disso, foram definidas cinco tarefas principais durante a análise das entrevistas. Elas são: aquisição de conhecimento, planejamento do documento, composição, validação e manutenção.

A tarefa de aquisição de conhecimento é muito difícil e ocupa tanto tempo quanto a própria escrita dos documentos técnicos. Os autores têm que trabalhar junto com os engenheiros e designers para obter informações sobre a estrutura e os procedimentos do produto, consultam o código fonte dos programas para ter acesso aos comentários e também experimentam protótipos do produto. No primeiro momento que os autores entram em contato com o produto, eles criam um modelo mental do mesmo, da perspectiva do usuário. Um registro formal deste modelo seria útil na documentação de subseqüentes mudanças do produto pelo mesmo autor ou por colegas. Isto poderia ser usado quando um mesmo procedimento necessita ser explicado novamente em um contexto diferente. Apesar de importante, esse modelo não é criado explicitamente.

No planejamento do documento, os autores devem estabelecer a estrutura geral e o objetivo do documento. A estrutura típica do um manual de instruções é ter um pequeno capítulo para cada tarefa, dividido em operações com cinco ou seis linhas.

Na composição, geralmente são feitos vários rascunhos para um documento. Os autores devem ser comunicadores eficientes, evitando jargões e transmitindo a informação com termos claros e concisos. Com treinamento e experiência, os autores se familiarizam com as normas padrões de redação técnica. A maioria das companhias também possui guias de estilo, que podem ser formais ou informais; podem, por exemplo, recomendar ou prescrever o uso de construções e terminologias específicas. Um guia de estilo rigorosamente formalizado impõe uma linguagem controlada, que pode ser difícil e demorada para se dominar.

A validação é feita por mecanismos de controle de qualidade que vão desde a leitura informal por colegas até a revisão formal feita por comitês. Algumas organizações passam os documentos finais por verificadores automáticos de gramática e terminologia.

A manutenção dos documentos é feita sempre quando um produto sofre alguma alteração. Uma grande parte do tempo dos autores é destinada a essa tarefa.

Após a análise das entrevistas e do entendimento da tarefa, foram definidas as funções desejáveis para uma ferramenta de suporte à escrita. São elas:

- Suporte para reuso de conhecimento através do auxílio da criação de um modelo de conhecimento adquirido pelos autores – Essa função seria útil para formalizar e organizar de maneira consistente o conhecimento adquirido, permitindo assim seu reuso e compartilhamento com colegas de trabalho.
- Produção de formulações alternativas quando possível – Os autores desejam a possibilidade de ter opções na hora de expressar um conjunto de instruções, pois existem vários caminhos para isso, e eles querem escolher a mais apropriada.
- Disponibilização dos primeiros rascunhos (daí o nome da ferramenta – DRAFTER) produzidos simultaneamente em várias línguas – Quanto mais rápido os rascunhos forem disponibilizados, mais rápido serão descobertas quais conhecimentos estão faltando pra determinada língua, além de agilizar todo o processo.
- Propagação de alterações por todo o documento e línguas – Ao efetuar uma alteração no texto, as demais partes do documento que contêm a mesma informação também devem ser atualizadas.
- Suporte para terminologia consistente e precisa – termos técnicos devem ser empregados consistentemente dentro e entre documentos, mesmo se forem produzidos por autores diferentes.
- Manter a satisfação criativa da escrita técnica – uma ferramenta de suporte a escrita deve **automatizar** aspectos que os escritores acham tediosos, como por exemplo, revisão e alguns aspectos rudimentares da composição (terminologia e sintaxe), deixando o autor livre para trabalhar melhor na estruturação do conteúdo e transmissão das idéias.

Baseado na análise de requisitos que foi descrita até agora, foi planejado e implementado o DRAFTER. A arquitetura da ferramenta é mostrada na Figura 2.

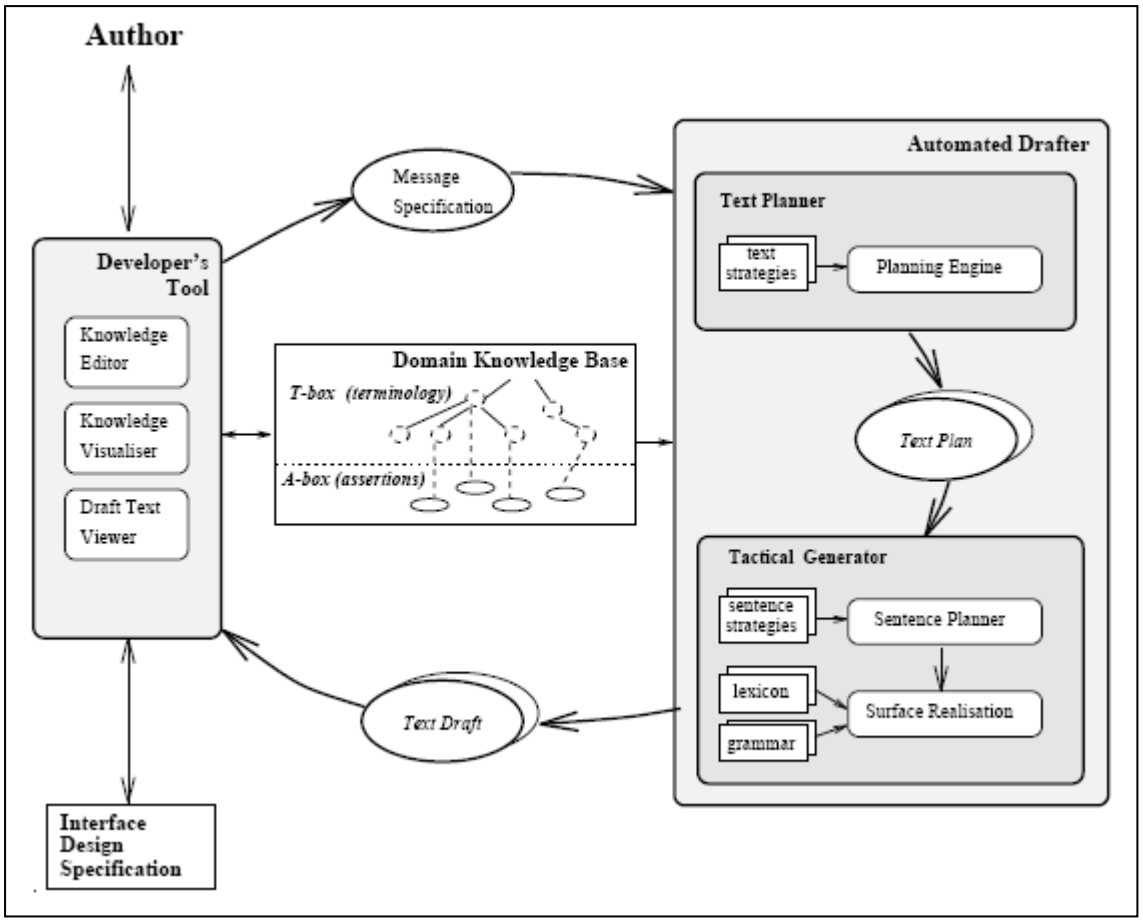

Figura 2 - Diagrama de Blocos da Arquitetura de DRAFTER (Paris *et al*., 1995)

O DRAFTER contém três módulos de processamento, que formam duas ferramentas de suportes principais:

- **Uma interface para o escritor técnico**. Permite ao autor especificar formalmente os procedimentos necessários para o usuário atingir seu objetivo. Ajuda também os escritores a controlar o processo de elaboração.
- **A ferramenta de elaboração**. Compreende dois componentes principais: o planejador estratégico e o gerador tático. O planejador estratégico determina o conteúdo e a estrutura do texto, e o gerador tático executará a realização das sentenças. O resultado será **rascunhos** das instruções em Inglês e Francês definidos pelo autor através da interface.

Dando base aos componentes de processamento, existe um modelo do domínio, que é o repositório principal de informações sobre o domínio.

O Modelo do Domínio é uma coleção de entidades representando a informação comumente usada no domínio de software. Essas entidades incluem ações, estados, objetos e

uma série de relações entre elas. Este conhecimento é derivado de um **estudo de um corpus multilíngüe de manuais de software**, e é tratado como independente de língua, que é um requisito básico para geração multilíngüe. Os autores usam os conceitos e relações do modelo de domínio para especificar os procedimentos apropriados para o sistema de software em particular que está sendo documentado. Este estudo de corpus é de fundamental importância para este mestrado e será descrito na Seção 2.2.2.1.

O DRAFTER fornece uma interface para criar e manter um registro formal do conhecimento que o autor aprendeu durante a tarefa de aquisição de conhecimento. Ela permite aos usuários especificar o conhecimento conceitual requerido para a tarefa a ser documentada, e informação importante do ponto de vista do usuário. Ela explora o modelo de domínio para fornecer guia e estrutura. A interface engloba as seguintes funções:

- Construção e manutenção da base de conhecimento de afirmações que contém a descrição das funções que o usuário pode realizar com o software que está sendo documentado.
- Visualização dos aspectos da base de conhecimento.
- Visualização e edição dos rascunhos gerados automaticamente.

Todas essas funções são chamadas através de menus, ícones e outros objetos sensíveis ao mouse. A base de conhecimento de afirmações pode ser atualizada com a inclusão de novas informações. Essas alterações são feitas através de um editor de conhecimento. Através desse editor, o autor escolhe um nó apropriado no modelo de domínio no qual a nova informação será subordinada. Para isso, são usados menus para escolher o conceito a ser atualizado. Após a escolha do conceito, o sistema gera um formulário dinamicamente para que o autor entre com as informações. O sistema também possibilita a anotação de imagens com informação semântica.

O DRAFTER possui também um visualizador da base de conhecimento, que possibilita os autores consultar relacionamentos entre ações, métodos e sub-ações, através de uma forma gráfica. Além disso, o visualizador é integrado com as funções de construção e edição da base de conhecimento.

Na visualização e edição dos rascunhos que foram gerados automaticamente, o texto é sensível ao mouse, permitindo ao autor consultar a base de conhecimento a partir da parte do texto que foi selecionado. Também é possível fazer alterações no texto que foi gerado automaticamente, através de um editor que também é disponibilizado.

O DRAFTER usa um sistema de planejamento de texto que monta uma árvore de discurso através de um objetivo comunicativo que foi informado pelo autor. A partir desse objetivo, o sistema busca em sua biblioteca de estratégias de discursos um plano capaz de atingir tal objetivo. Nessa árvore, o nó raiz representa o objetivo inicial que foi traçado pelo plano, e as folhas representam os passos a serem tomados. Essa árvore também inclui relações de coerência indicando como as várias partes dos textos são retoricamente relacionadas. Algumas limitações impostas pelos padrões da escrita ou pelo estilo da empresa podem ser unidas à estratégia de discurso, e pela associação de várias estratégias para o mesmo objetivo, é possível construir rascunhos alternativos, conforme desejado pelo autor.

A cobertura do gerador tático no Inglês foi estendida para gerar os tipos de sentenças encontrados em manuais de instruções, e seu ambiente flexível foi usado para desenvolver uma gramática em Francês.

2.2.2.1 A Análise de Corpus no Projeto DRAFTER

Nesta seção, descrevemos o trabalho de Paris and Scott (1994) que apresenta um estudo sobre a variação estilística dos manuais de instrução. A pesquisa mostra que os manuais podem ter diferentes estilos, por exemplo, nem todas as instruções são registradas usando uma seqüência de imperativos. Além disso, diferentes partes dos manuais usam diferentes estruturas de discurso e formas de realização. Este trabalho deu suporte às decisões de geração multilíngüe no projeto DRAFTER, descrito na Seção 2.2.2 Geração Multilíngüe de Manuais de Instrução. Também, relatamos as freqüências das formas gramaticais para expressar as relações gera e habilita, encontradas no corpus de instruções em português do estudo de Delin *et al*. (1994), por ser o português a língua de interesse deste trabalho de mestrado.

2.2.2.1.1 Variação Estilística dos Manuais de Instrução

Instruções são usadas para transmitir direções para se cumprir uma tarefa. Entre essas direções, estão incluídas ações que devem ser realizadas sobre instalações, manutenção e também avisos e alertas sobre a segurança da execução das tarefas. O modo mais direto de fazer com que o leitor execute uma tarefa é o uso de seqüências de imperativos. Em Paris & Scott (1994) foi verificada, na análise do corpus, que existe uma variação na transmissão das instruções, mostrando que em alguns casos são usadas declarações simples. Muitos podem considerar que declarações simples estão fora do escopo de instruções, mas o estudo também mostra que estas declarações são freqüentemente instruções implícitas.

Essas variações nos manuais de instruções estão ligadas à postura tomada pelo autor em relação ao leitor. Geralmente, diferentes posturas são adotadas na descrição de produtos e também em avisos e advertências. Essas posturas são geralmente uma decisão institucional do fabricante do produto, e são chamadas de "estilo da casa". Elas são usadas para projetar uma imagem da empresa para os leitores.

Na análise do corpus de 30 manuais de produtos em Inglês e Francês que tinham entre 1 a 20 páginas tentou-se identificar: as diferentes atitudes que os textos expressavam e as realizações. Na análise, foram identificados quatro tipos diferentes de postura que um manual pode adotar, com exemplos em francês e inglês mostrados na Figura 3, que são apresentados abaixo:

- **Fornecimento de Informação**: O texto concentra-se na informação factual a ser transmitida. Essa postura pretende aumentar o conhecimento do leitor sobre o produto/tarefa em questão. Para transmitir informação, os escritores mostram forte preferência pelo uso de declarações ativas simples.
- **Elogio**: O texto enfatiza os aspectos positivos do produto e parabeniza o leitor pela boa escolha na aquisição. Neste caso também são usadas declarações simples ativas, mas muito qualificadores são usados, geralmente são adjetivos positivos e superlativos. Eles são usados para mostrar a superioridade dos atributos do produto, o conforto, a qualidade do material, entre outros.
- **Diretivas sobre como realizar uma tarefa**: O leitor deve realizar uma tarefa exatamente como prescrita. A lógica por trás das prescrições não é considerada necessária. A função principal do manual de instruções é fazer com que o leitor execute ou evite ações específicas. O jeito mais simples de fazer isso é através do fornecimento de direções que o usuário deve tomar para fazer ou não alguma ação. A preocupação é que o leitor siga as instruções, e não que ele saiba o porquê da ação ser ou não executada. O texto leva a autoridade do autor sobre o leitor e o deixa sem escolha. São usados atos da fala de diretivas fortes como ordens e proibição. Explicações sobre essas diretivas raramente são dadas. A forma como essas diretivas são passadas para o leitor podem variar conforme o quanto o autor deseja ser pessoal/impessoal. Essas diretivas podem ser passadas através do uso de imperativos, quando o autor está bem "próximo" (referência direta) do leitor. Quando o autor não está tão próximo assim do leitor, o uso de "deve" na forma modal, acompanhado de um advérbio pode ser usado. Quando o autor está muito

distante do leitor (a ordem não está referida diretamente ao leitor), as diretivas são dadas na forma passiva, ou através de verbos que requerem que o objeto se torne o sujeito sintático, junto com um advérbio para enfatizar. Em todos os casos, diretivas fortes podem ser transmitidas usando termos como "nunca" ou "sob nenhuma circunstância" como parte da sentença ou mesmo como uma sentença separada.

- **Explicações sobre a melhor maneira de completar uma tarefa**: O leitor recebe ajuda sobre como alcançar uma tarefa, e uma explicação do porquê deveria ser feito da maneira prescrita. Isto oferece uma oportunidade para o leitor de construir um modelo mental da tarefa ou do produto.

(1) Information Provision – from Dietrich self-cleaning enamel oven:
L'émail auto-nettoyant est de couleur brun foncé moucheté de blanc pour certaines pièces.
Loose English Translation: *The self-cleaning enamel is dark brown, speckled with white on some parts.*

---

(2) Information Provision – from HP LaserJet 4 Printer Installation Guide:
The optional 500-sheet paper tray assembly comes equipped with a tray housing and either a legal, letter, A4, and executive multi-size paper tray or a letter, A4, and executive multi-size paper tray. You may purchase replacement trays from your authorized HP dealer. The part number of the letter, A4 and executive multi-size paper tray is C2084B. . .

---

(3) Eulogy – from Sennheisser Headphones:
To wish to convince you of the superior quality of the dynamic open air headphone HD 40 would be something of a paradox as you are by now already in possession of this product. However, the arguments for the HD 40 are in fact very convincing:

- High quality reproduction.
- Extremely comfortable thanks to very low weight.
- Problem-free connection by means of universal connector.
- Very flat storage space thanks to turntable driver elements.

---

(4) Directive – from Krupp Expresso Coffee Machine:
Do not use or put down the appliance on any hot surface (such as a stove hot-plate) or in the vicinity of an open gas flame.

---

(5) Directive – from Camping Stove: Camping Gaz International:
Utilisez toujours votre réchaud dans un endroit suffisamment aéré, sans l'envelopper dans un objet quelconque et pas trop près de matières inflammables.

Loose English Translation: *Always use your stove in a well ventilated place, without wrapping it in any object and not too close to flammable substances.*

---

(6) Explanation – from Dietrich Oven:
To avoid heavy spittings on the sole plate, it is advisable to cover it with an aluminium foil.

---

(7) Explanation – from EMSA Thermos
Attention: Pour éviter d'abimer le récipient en verre, ne remuez pas les boissons avec des cuillères métalliques.
Loose English Translation: *Warning: To avoid damage to the glass jug, do not stir drinks with metallic spoons.*

Figura 3 - Exemplos de posturas que um manual de produtos pode adotar (Paris & Scott, 1994)

Através da análise do corpus, também foi definida a estrutura global dos manuais de instruções e quais as posturas adotadas em cada parte. O resultado da análise traz uma estrutura formada por seis partes principais, apresentada na Figura 4.

Informações gerais

Esta parte do manual contém agradecimentos pela compra do produto, uma descrição do produto com suas vantagens e uma lista de partes e também apresenta as condições de garantias (ou a falta delas). Nesta parte do manual, as posturas utilizadas são a de *Fornecimento de informação e elogios***.**

Informações sobre segurança

Esta parte inclui avisos, conselhos gerais de segurança e passos cruciais que devem ser realizados, sejam eles para realizar a tarefa ou para obter melhores resultados. A postura tomada nesta parte pode ser o uso de *diretivas ou de explicações***.**

Etapa de preparação ou instalação

Contém informações sobre como instalar o aparelho, ou sobre quais passos devem ser realizados antes do primeiro uso. O uso de *diretivas e explicações* é apropriado nesta seção.

Uso

Contém os passos ou ações que devem ser realizadas pelo leitor para usar o produto e atingir os objetivos de diferentes tarefas. O uso de *diretivas e explicações* é apropriado nesta seção.

Limpeza, Cuidado e Manutenção

Nesta seção o autor apresenta para o leitor como ele deve limpar e cuidar do produto. O uso de *diretivas e explicações* é apropriado nesta seção.

Resolução de problemas

O objetivo desta parte do manual é ajudar o leitor a identificar a fonte do problema e fornecer informação sobre o que poderia acontecer de errado se um passo não for propriamente realizado. Ações a serem tomadas, junto com condições sobre as quais são apropriadas e soluções para os problemas, são fornecidas. A postura utilizada nessa seção é, geralmente, a *diretiva***.**

Figura 4 - Estrutura Global de Manuais de Produtos (Paris & Scott, 1994)

Alguns manuais contêm todas as seis partes da Figura 4, claramente indicadas. Entretanto, há manuais que, devido a problemas de espaço, principalmente, intercalam as partes. Um exemplo de intercalação de partes encontrada na análise de Paris & Scott (1994) foi a apresentação de cuidados logo após uma tarefa ser expressa no manual. É importante notar que uma estrutura global, também chamada de estrutura esquemática, apresenta suas partes/componentes numa ordem esperada e, portanto, tal ordem das partes não pode aparecer aleatoriamente.

2.2.2.1.1 Realização das Expressões Gera e Habilita em Português

O trabalho de Delin *et al*. (1994) trata de textos originais nas línguas inglesa, francesa e portuguesa. Os resultados do estudo serviram de base para o projeto DRAFTER, pois informam quais são os mapeamentos mais comuns das relações gera e habilita para uma língua e se há ambigüidade entre as formas da língua para estas relações.

O corpus do português do Brasil foi formado por 8 manuais de instrução, que possuem de 2 a 7 páginas. Os manuais são de produtos e são dirigidos para o público não-especialista, que é o mesmo público que nos dedicamos neste projeto de mestrado. A avaliação de freqüência se deu em 33 relações gera e 32 habilita. A maioria das relações gera envolvem o uso do conectivo "para" seguido de um infinitivo, como pode ser observado na Tabela 1.

Tabela 1 – Freqüência das formas gramaticais em português para a relação gera (Delin *et al*., 1994).

Forma gramatical	Contagem	Porcentagem
Para + Infinitivo	23	70%
Se + Subjuntivo	3	9%
Para + Frases	2	6%
Seqüência	1	3%
2 outras formas	4	12%
Total:	33	

Para relações habilita, o uso de seqüências (Uma seqüência de imperativos, tanto em frases consecutivas, ou dentro de frases e ligadas pela conjunção "e") é predominante, como mostrado na Tabela 2.

Tabela 2 - Frequência das formas gramaticais em português para a relação habilita (Delin *et al*., 1994).

Forma gramatical	Contagem	Porcentagem
Seqüência	25	78%
Condição "depois"	4	12%
Condição "antes"	3	9%
Total:	32	

2.2.3 Estrutura de Textos Instrucionais e Sistemas de Perguntas & Respostas

No trabalho de Farida Aouladomar (Aouladomar, 2005a; 2005b; Aouladomar & Saint-Dizier, 2005) foi feita uma análise da estrutura de perguntas procedimentais "Como?" e "Por quê?".

Também foi feita uma pesquisa sobre a estrutura de textos procedimentais e as relações retóricas relevantes para responder as perguntas.

O trabalho é baseado na língua Francesa e utiliza estudos baseados em corpus de sistemas de pergunta e resposta procedimentais de vários domínios, entre eles estão saúde, educação, turismo, comportamento social, ciência da computação e manutenção.

As perguntas procedimentais geralmente são introduzidas pelo pronome interrogativo "*Como"*. No entanto, existem vários casos em que esse pronome não está relacionado com perguntas procedimentais, como por exemplo:

- Situações nominais: Como isto se chama?
- Situações causais: Como isto aconteceu?
- Situações booleanas (Verdadeiro ou Falso): Como você está?
- Usado como forma instrumental: Como cuscuz é comido em Marrocos?
- Situação de escolha: Como posso pagar o ingresso?

A única situação nesta análise em quem o pronome "Como" é usado em perguntas procedimentais é a situação instrumental:

Como posso trocar o pneu do meu carro?

Existem outras formas de iniciar uma pergunta procedimental:

- Na forma de "*O que fazer para...*": O que fazer para conseguir um visto para os Estados Unidos?
- Perguntas usando (*"É possível" ou "Posso")* + Sintagma verbal: É possível criar diretórios usando php?

Existem quatro categorias conceituais para perguntas procedimentais que foram encontradas em 90% dos casos do corpus de estudo. Para terem uma análise que é simples e fácil de adaptar para outras línguas, foram consideradas as categorias básicas de verbos, como as utilizadas na WordNet (Fellbaum, 1998) que foram adaptadas para o Francês em (Saint-Dizier, 1998). As categorias são apresentadas a seguir:

- **Ação**: Caracterizada pelo uso de verbos de mudança, criação e destruição e manutenção: *construir, revisar, montar, desmontar, reparar, mudar*. Aparecem em textos procedimentais de receitas, guias "Faça você mesmo", manuais de manutenção e construção.

- **Comunicação**: Caracterizada pelos verbos de interação social e verbos psicológicos: *Contatar, negociar, convencer, gerencia*. Estes aparecem em horóscopos, guias de gerenciamento e educação.
- **Aquisição de conhecimento**: Caracterizada principalmente por verbos de cognição e verbos que expressam formas de transferência de conhecimento: *Aprender, resolver, melhorar*. Estes tipos de textos são encontrados em enciclopédias e textos didáticos.
- **Itinerário**: Caracterizada por verbos de movimento: *ir, alcançar, acessar*. Aparecem em guias de viagem e de conduta.

A representação das perguntas procedimentais deste trabalho, utilizada em experimentos com perguntas procedimentais curtas, possui o formato:

Pergunta (procedimento (tipo), foco, restrições).

Tipo é um dos quatro tipos que foram apresentados anteriormente, foco é geralmente o sintagma verbal: o predicado e seus argumentos, que caracterizam o objetivo e as restrições são compostas por adjuntos do predicado. Como exemplo, a pergunta:

Como reservar um vôo na web?

É representado da seguinte maneira:

Pergunta (procedimento (ação), [reservar (verbo, morfológico), vôo (substantivo, transporte)], [na (preposição, meio)], web (substantivo, comunicação)])

A estrutura de textos procedimentais foi elaborada na forma de uma gramática. Nesta gramática, parênteses representam opção, + significa iteração, a vírgula é usada apenas com um separador, / significa "ou" e o operador < indica a precedência preferida. Estes símbolos correspondem a etiquetas em XML, e permitem a anotação dos textos instrucionais.

O nó topo é denominado objetivo:

Objetivo -> título, (sumário), (advertência), (pré-requisitos), (figura)+ < seqüências de instrução.

Sumário -> título+. Descreve a organização global do procedimento. Útil quando o procedimento é complexo.

Advertência -> texto, (figura)+, (pré-requisitos). Representam precauções globais e prevenções associadas com ações ou objetivos.

Pré-requisitos -> lista de objetivos, seqüência de instruções. Descreve todo tipo de equipamento necessário para executar determinada tarefa.

Figuras servem para demonstrar esquemas de montagens e diagramas de vários tipos. São usadas para tornar as instruções mais claras.

Seqüências de instruções são estruturadas da seguinte forma:

Seqüências de instruções -> seqint < conectores discursivos < seqüências de instruções / seqint.

Seqint é um dos quatro tipos abaixo:

Seqint -> (objetivo), seqüência imperativa linear / (objetivo), seqüência opcional / (objetivo), seqüência alternativa / (objetivo), seqüência imperativa co-temporal.

Cada tipo de seqüência de instrução é definido a seguir:

Seqüência imperativa linear -> instrução < (marca temporal), sequência imperativa linear / instrução.

Seqüência opcional -> expressão condicional, seqüência imperativa linear.

Sequência alternativa - > (expressão condicional), seqüência imperativa linear, (marca de oposição alternativa) < seqint / (expressão condicional, seqint)+

Seqüência imperativa co-temporal -> seqüência imperativa linear < marca co-temporal < seqüência imperativa co-temporal / instrução.

Finalmente, a instrução é o último nível e possui recursão no objetivo:

Instrução -> (expressão iterativa), ação, (objetivo)+, (referência)+, (maneira)+, (motivação), (limite), (figura)+, (advertência) / objetivo.

A estrutura retórica desempenha vários papéis nesta abordagem. Ela contribui para reforçar a produção de respostas bem formadas, permite a integração entre textos procedimentais com objetivos similares, apesar de ser uma tarefa difícil, e também é usada para responder perguntas com uma alta exatidão por identificar claramente instrumentos, riscos e equipamentos necessários. Foram identificadas 17 relações retóricas na análise do corpus, que serão apresentadas a seguir:

- **Seqüência:** é uma relação multi-núcleo onde os núcleos são ligados por uma relação de sucessão.
- **Resultado:** especifica que uma ação não pode começar antes que o resultado desejado de uma ação anterior seja alcançado.

- **Propósito:** ocorre entre o objetivo e as ações destinadas a alcançar tal objetivo.
- **Avaliação:** é uma pequena diferença entre o propósito e a relação de resultado por que é possível avaliar quando a ação foi feita corretamente ou não.
- **Limite:** Liga uma ação com um satélite que representa uma interrupção.
- **Alternativa:** liga duas ações alternativas.
- **Meios:** É a ligação entre uma ação e os instrumentos necessários para executar tal ação.
- **Referência:** fica entre uma ação e um seguimento que fornece a localização de um procedimento no texto ou em textos relacionados através de um hyperlink.
- **Pré-requisitos:** ocorre entre um objetivo ou ação e uma lista de instrumentos sem os quais a ação ou objetivo não é realizado.
- **Opção:** é considerada quando uma ação depende da realização de uma situação condicional.
- **Prevenção:** Relação normalmente encontrada entre uma ação e uma advertência.
- **Condição:** aparece quando a ação resulta da ocorrência de uma situação condicional.
- **Co-ocorrência:** ocorre quando o núcleo é ligado por uma relação co-temporal.
- **Concorrência:** ocorre entre duas ações co-temporais rivais.
- **Motivação:** ocorre quando a informação enviada pelo satélite faz com que o leitor aumente a vontade de executar a ação.

3 Simplificação Sintática no PorSimples

O projeto PorSimples (Aluísio *et al*., 2008a; 2008b)(*Simplificação Textual do Português para Inclusão e Acessibilidade Digital*), aprovado no âmbito do Edital Microsoft-Fapesp (proc. nro. 2007/54565-8), aborda a tarefa de simplificação textual com vistas à construção de sistemas para promover o acesso de textos escritos em português por pessoas com níveis de letramento rudimentares e básicos, com problemas cognitivos como afasia e dislexia, além de adultos e crianças em fase de aprendizado da leitura e escrita. Neste projeto, estão sendo desenvolvidos dois tipos de sistemas, (i) um de auxílio à edição de textos que serão colocados na Web e o outro (ii) que permite a leitura de textos já publicados através de uma versão simplificada produzida por ele (Aluísio *et al*., 2008a). Este último é composto de três módulos: sumarização, simplificação e explicitação visual de certas características do texto, visando uma facilitação da leitura.

Para a construção dos sistemas de simplificação, duas abordagens estão sendo exploradas. A primeira abordagem é simbólica e constrói sistemas de simplificação via regras desenvolvidas manualmente, com ajuda de informação sintática e também discursiva. A segunda abordagem explora a indução destas regras a partir de corpus alinhados de sentenças originais e suas correspondentes simplificadas. Esta última abordagem faz uso de métodos estatísticos que precisam de um grande volume de textos anotados manualmente.

Para explorar a abordagem simbólica, um manual de simplificação sintática foi elaborado, com base na *Plain Language*[16], em trabalhos de geração e simplificação automática para língua inglesa (Siddharthan, 2002; Williams, 2004), na análise de textos simples disponíveis na Web brasileira. Exemplos de textos simples da Web são a cartilha **Brasileiras e Brasileiros no Exterior**[17], que é uma cartilha criada pelo governo que utiliza uma linguagem simples e acessível a pessoas de qualquer idade e com um menor grau de instrução e bulas de alguns remédios do **Bulário da Anvisa**[18], pois o governo editou leis que tornam as bulas de medicamentos mais fáceis de serem lidas, separando as informações para o público geral da informação que é voltada para os médicos e farmacêuticos e nos fenômenos da gramática do português que tornam um texto complexo (Specia *et al*., 2008).

As operações de simplificação do manual, que visam tornar as sentenças mais curtas, simples, com apenas uma proposição e que sigam a ordem Sujeito-Verbo-Objeto (S-V-O),

[16] http://www.plainlanguage.gov/
[17] http://download.uol.com.br/ultnot/cartilha.pdf
[18] http://bulario.bvs.br/

serão apresentadas a seguir (Specia *et al*., 2008).Além de servir de base para a criação do método de simplificação simbólico, o manual de simplificação serviu como guia para anotadores humanos simplificarem textos e criarem um corpus de textos simplificados anotados com as operações efetuadas. São seis os fenômenos lingüísticos tratados no manual: apostos, cláusulas relativas, cláusulas subordinadas, coordenadas e sentenças na voz passiva.

1 – Aposto (ou orações subordinadas substantivas apositivas)

A – Enumerativo

Regra: Não simplificar. É usada, geralmente, com recursos para-lingüísticos de itemização e enumeração.

B – Recapitulativo

Regra: Não simplificar.

C – Especificador

Regra: Dividir a sentença em duas:

1) Sentença original sem o aposto
2) Sentença para o aposto, a qual terá como sujeito o núcleo do termo da oração principal explicado pelo aposto (ou o termo completo, caso o núcleo não seja suficiente), seguido de verbo de ligação ‘ser’ no tempo presente e então do aposto.

2 – Orações subordinadas adjetivas (ou orações relativas)

A – Explicativas: não limitam o sentido do antecedente, acrescentam uma informação que pode ser eliminada sem prejuízo para a compreensão lógica da sentença. Vêm entre vírgulas, sempre introduzidas por pronome relativo (que, quem, o qual, a qual, os quais, as quais, onde, cujo, quanto, etc.)

Regra: Dividir a sentença em duas:

1) Sentença original sem a cláusula subordinada.
2) Sentença para a cláusula subordinada, a qual terá como sujeito o núcleo do termo da oração principal explicado pela subordinada (ou o termo completo, caso o núcleo não seja suficiente), seguido da subordinada, propriamente dita, eliminando-se seu pronome relativo.

Regra para 'o que': Dividir a sentença em duas:

1) Sentença original sem a cláusula subordinada.
2) Sentença para a cláusula subordinada, a qual terá como sujeito 'isso', seguido da subordinada, propriamente dita, eliminando-se o pronome relativo 'o que'.

B – Restritivas: restringem, limitam a significação do seu antecedente (substantivo ou pronome). São indispensáveis ao significado da sentença e não se separam da oração principal por vírgula.

Regra: Dividir a sentença em duas:

1) Sentença para a cláusula subordinada, a qual terá como sujeito o núcleo do termo da oração principal restringido pela subordinada (ou o termo completo, caso o núcleo não seja suficiente), seguido da subordinada, propriamente dita, eliminando-se seu pronome relativo.
2) Sentença original sem a cláusula subordinada.

3 – Orações subordinadas adverbiais

A – Causais: indicam a causa da ação expressa pelo verbo na oração principal: pois, que, uma vez que, visto que, visto como, porquanto, porque, como, já que, etc.

Regra: Dividir a sentença em duas:

1) Sentença para a cláusula subordinada, a qual terá como sujeito o núcleo do termo da oração principal modificado pela subordinada (ou o termo completo, caso o núcleo não seja suficiente), seguido da subordinada, propriamente dita, eliminando-se o marcador discursivo.
2) Sentença original sem a cláusula subordinada, introduzida por um marcador discursivo que indique efeito, preferencialmente, 'COM ISSO'.
   ➔ Manter a ordem canônica: causa -> efeito.

B – Comparativas: estabelecem uma comparação com a ação indicada pelo verbo na oração principal: tão que/do que, tanto que/do que, pior que/do que, melhor que/do que, maior que/do que, como, assim como, assim, etc.

Regra para 'assim como', 'tanto...quanto' e similares: Dividir a sentença em duas:

1) Sentença original sem a cláusula subordinada.
2) Sentença para a cláusula subordinada, eliminando-se o marcador de comparação e acrescentando-se 'também', seguido pela reprodução do verbo da oração principal, caso este esteja oculto. O sujeito da sentença pode precisar ser reproduzido.
   ➔ Definir a posição do marcador 'também' procurando manter a ordem S-V-O.

C – Concessivas: indicam a concessão à idéia expressa pelo verbo da principal, isto é, admitem uma contradição ou um fato inesperado: embora, por mais que, ainda que, posto que, a menos que, se bem que, conquanto, mesmo que, nem que, apesar de que, por mais que, por muito que, etc.

Regra: Dividir a sentença em duas:

1) Sentença para a cláusula subordinada, eliminando-se o marcador discursivo de concessão e eventualmente introduzindo-se como sujeito o núcleo do termo da oração principal a quem a subordinada se refere (ou o termo completo, caso o núcleo não seja suficiente). O tempo verbal pode precisar ser adaptado.
2) Sentença original sem a cláusula subordinada, introduzida pelo marcador discursivo 'MAS'.
   Ordem: fato1, embora fato2 VIRA fato 2 mas fato1

Regra para orações que dão idéia de que a parte subordinada é hipotética (por exemplo, com "ainda que / por mais que"): Dividir a sentença em duas:

1) Sentença original sem a cláusula subordinada.
2) Sentença para a cláusula subordinada, eliminando-se o marcador discursivo de concessão e introduzindo-se 'ISSO ACONTECE MESMO QUE'.

D – Condicionais: indicam a situação necessária à ocorrência ou não da ação do verbo da principal: se, sem que, caso, exceto, salvo, desde que, contanto que, a menos que, a não ser que, etc.

Regra: Não DIVIDIR.

➔ Manter a ordem: condição -> efeito.

Regra para discurso indireto: Não dividir a sentença.
Manter a ordem: condição -> efeito.

Regra para condição com negação (por exemplo, marcadores 'a menos que', 'sem que', etc.): Não dividir a sentença, iniciá-la com a condição **negada** (precedida pelo marcador 'SE') e finalizada com o efeito (precedido pelo marcador 'ENTÃO').

➔ Manter a ordem: condição -> efeito.

E – Consecutivas: indicam a conseqüência resultante do fato expresso pelo verbo da oração principal: tão...que, tanto...que, tal...que, ...que, de forma que, de modo que, de sorte que, tanto que, tamanho....que, etc.
Regra: Dividir a sentença em duas:

1) Sentença original sem a cláusula subordinada.
2) Sentença para a cláusula subordinada, eliminando-se o marcador discursivo de conseqüência e introduzindo-se o marcador 'POR ISSO,' no início da sentença e, eventualmente, o sujeito constituído do núcleo do termo da oração principal a que se refere (ou o termo completo, caso o núcleo não seja suficiente). O tempo verbal pode precisar ser alterado.

F – Finais: indicam o fim, o objetivo do fato enunciado na oração principal: para que, a fim de que, porque, etc.
Regra: Dividir a sentença em duas:

1) Sentença original sem a cláusula subordinada.
2) Sentença para a cláusula subordinada, eliminando-se o marcador discursivo de finalidade e introduzindo-se o marcador 'O OBJETIVO É QUE,' no início da sentença.

G – Proporcionais: indicam uma relação de proporcionalidade com o verbo da oração principal: à proporção que, à medida que, ao passo que, quanto mais... mais, quanto mais... menos, etc.
Regra: Não simplificar.

H – Conformativas: indicam uma conformidade, um acordo entre o fato que expressam e a ação do verbo da oração principal: conforme, como, consoante, segundo...

Regra: Inverter a ordem das cláusulas, considerando-se primeiramente a cláusula subordinada, com a explicitação do sujeito e a eliminação ou modificação do marcador discursivo de conformidade para a introdução de uma cláusula relativa precedida por 'que'. Essa cláusula relativa consiste da oração original sem a cláusula subordinada. Se não houver verbo não oração subordinada, deve-se utilizar 'CONFIRMA QUE' ao invés de 'QUE'.

➔ Procurar manter a ordem S-V-O

I – Temporais: indicam a circunstância de tempo em que ocorre a ação do verbo da oração principal: quando, enquanto, assim que, logo que, até que, depois de, desde que, apenas, mal, sempre que, cada vez que, antes que, etc.

Regra: Dividir a sentença em duas:

1) Sentença para a cláusula subordinada, eliminando-se o marcador discursivo e adequando-se o tempo verbal se necessário. Pode ser necessário incluir o sujeito constituído do núcleo do termo da oração principal a que se refere a subordinada (ou o termo completo, caso o núcleo não seja suficiente).
2) Sentença original sem a cláusula subordinada. O tempo verbal pode precisar ser alterado.

Para marcadores que indicam seqüência no tempo (ex.: depois que, logo que, etc.): Dividir a sentença em duas:

1) Sentença para a cláusula subordinada, eliminando-se o marcador discursivo e adequando-se o tempo verbal se necessário. Pode ser necessário incluir o sujeito constituído do núcleo do termo da oração principal a que se refere a subordinada (ou o termo completo, caso o núcleo não seja suficiente).
2) Sentença original sem a cláusula subordinada, precedida pelo marcador 'EM SEGUIDA'. O tempo verbal pode precisar ser alterado.

4 – Orações coordenadas

A – Assindéticas

Regra: Dividir a sentença em tantos quantos forem os períodos, eliminando-se o delimitador de período e repetindo-se o núcleo do sujeito e o verbo (se oculto) para cada período, quando necessário.

B – Sindéticas

1º. Aditivas: expressam uma adição, uma seqüência de informações; e, nem, não só, mas também, ...

Regra: Dividir a sentença em tantos quantos forem os períodos, eliminando-se o marcador de coordenação e repetindo-se o núcleo do sujeito e o verbo (se oculto) para cada período, quando necessário.

2º. Adversativas: expressam idéia de oposição, contraste: mas, porém, todavia, contudo, no entanto, entretanto, etc.

Regra: Dividir a sentença em duas:

1) Sentença original sem a cláusula subordinada.
2) Sentença para a cláusula subordinada, padronizando-se o marcador discursivo de oposição como 'MAS'. Pode ser necessário explicitar o sujeito, constituído do núcleo do termo da oração principal a que se refere a subordinada, bem como o verbo ou objeto (caso estejam ocultos).

3º. Alternativas: explicam alternância de idéias: ou...ou, ora...ora, já...já, quer...quer, etc.

Regra: Dividir a sentença em duas:

1) Sentença original sem a cláusula subordinada sem o marcador de alternância.
2) Sentença para a cláusula subordinada, excluindo-se o marcador de alternância e incluindo-se o marcador 'TAMBÉM' numa posição adequada, procurando manter a ordem S-V-O. Pode ser necessário explicitar o sujeito, constituído do núcleo do termo da oração principal, bem como o verbo ou objeto (caso estejam ocultos).

4º. Conclusivas: expressam a idéia de conclusão, conseqüência: logo, portanto, por conseguinte, pois (proposto ao verbo), etc.

Regra: Dividir a sentença em duas:

1) Sentença original sem a cláusula subordinada.

2) Sentença para a cláusula subordinada, excluindo-se o marcador de conclusão e incluindo-se o marcador 'COM ISSO'. Pode ser necessário explicitar o sujeito, constituído do núcleo do termo da oração principal, bem como o verbo ou objeto (caso estejam ocultos).

5º. Explicativas: indicam uma justificativa ou uma explicação ao fato expresso na primeira oração: porque, pois (anteposto ao verbo), que, etc.

Regra: Dividir a sentença em duas:

1) Sentença original sem a cláusula subordinada.
2) Sentença para a cláusula subordinada, excluindo-se o marcador de explicação e incluindo-se 'ISSO OCORRE PORQUE' no tempo verbal apropriado. Pode ser necessário explicitar o sujeito, constituído do núcleo do termo da oração principal, bem como o verbo ou objeto (caso estejam ocultos).

5 – Orações reduzidas

A – De infinitivo: não simplificar (ex.: era difícil andar)

B – De gerúndio: causal, concessiva, modal, condicional

Regra: Não simplificar.

C – De particípio: temporal, causal, concessiva, condicional.

Regra: Não simplificar.

6 – Orações em voz passiva

Regra: Transformar a sentença na voz ativa.

Usando as regras elaboradas no manual de simplificação (Specia *et al*., 2008) está sendo desenvolvido o SIMPLIFICA[19], que é uma ferramenta de auxílio à edição de textos que serão colocados na Web, citada no início deste capítulo. O SIMPLIFICA pode apoiar a escrita de textos destinados a analfabetos funcionais do nível básico e rudimentar. Para os primeiros, a decisão de quando simplificar ou não uma sentença é aprendida via métodos de aprendizado de máquina nos dados do corpus paralelo de textos originais e simplificados que pode ser

[19] http://www.nilc.icmc.usp.br/~erick/SS2/index.htm

pesquisado via um Portal de Corpora Paralelos de Simplificação[20]. Para os segundos, um sistema simbólico foi desenvolvido. Este simplifica toda sentença que possua um fenômeno lingüístico complexo, independente de outros fatores do texto. Nos dois casos, o sistema simbólico (regras) é utilizado: para simplificar a sentença escolhida ou para simplificar toda e qualquer sentença na qual se aplica uma das regras do manual.

Para aplicar as regras, é necessário ter conhecimento morfossintático e sintático do texto, e no caso do SIMPLIFICA foi utilizado o parser PALAVRAS (Bick, 2000), que é uma ferramenta que obtém um dos melhores resultados em análises morfossintáticas e sintáticas para o Português do Brasil. Após passar pelo parser, a ferramenta identifica quais regras devem ser aplicadas no texto e então faz as operações necessárias para realizar a simplificação e retorna o texto simplificado.

[20] http://caravelas.icmc.usp.br/portal/index.php

4 Métodos de Extração Automática de Termos

Devido ao grande volume de informação técnica disponível nos últimos anos, juntamente com a expansão da internet, a tarefa manual de extração de terminologia se tornou inviável.

Um sistema de extração automática de candidatos a termo (SEACAT) é formado por um conjunto de programas para o reconhecimento de unidades terminológicas de corpus (Estopà Bagot, 1999). Os candidatos a termos devem ser, posteriormente, validados por humanos. O objetivo dos SEACAT é a automatização da fase de seleção de todas as unidades terminológicas de um texto especializado.

Os SEACAT são classificados conforme a metodologia que utilizam para reconhecer as unidades terminológicas. São três as metodologias utilizadas (Teline, 2004):

- Metodologia baseada em conhecimento lingüístico;
- Metodologia baseada em conhecimento estatístico;
- Metodologia híbrida, baseada em conhecimento lingüístico e estatístico.

Essas metodologias serão detalhadas a seguir.

4.1 Métodos Lingüísticos

Os métodos baseados em conhecimento lingüístico utilizam diferentes recursos que fornecem diferentes informações lingüísticas, que são usadas para a extração dos termos:

- Lexicográficas: dicionários de termos e lista de palavras auxiliares (“stopwords”)
- Morfológicas: padrões de estrutura interna da palavra
- Morfossintática: categorias morfossintáticas e funções sintáticas
- Semânticas: classificações semânticas
- Pragmáticas: representações tipográficas e informações de disposição do termo no texto.

De acordo com Estopà Bagot (1999), a grande quantidade de ruído gerada (entre 55% e 75%) é um dos grandes problemas da abordagem lingüística. Além disso, ela é dependente da língua para qual foi desenvolvida, e também pode ser dependente de uma única variante. Dessa forma, a aplicação de um método lingüístico, desenvolvido para uma determinada língua deve ser reavaliado e adaptado para outra língua.

O trabalho de Heid *et al.* (1996) propõe um algoritmo com duas fases. A primeira fase consiste no pré-processamento lingüístico e anotação automática dos textos. Na segunda fase são feitas as consultas ao corpus e a extração de informações relevantes.

O trabalho de Klavans e Muresan (2000; 2001a; 2001b), que propõe um sistema baseado em regras que explora artigos orientados ao consumidor a fim de extrair definições e os termos que eles definem. Esta pesquisa pertencente ao Projeto de Biblioteca Digital na Universidade de Columbia. Uma meta desse projeto é apresentar definições precisas e legíveis de termos técnicos, que podem ser encontrados em artigos que apresentam complexidade intermediária, através do uso de uma linguagem que seja compreensível aos usuários. O algoritmo consiste na Combinação de técnicas superficiais de processamento de língua natural com análise gramatical, com o intuito de explorar textos eficientemente. É baseado em dois módulos:

1) Módulo de processamento de texto superficial que realiza análise de padrões com o uso de uma gramática de estado finito guiada por expressões lingüísticas, e um conjunto limitado de marcadores de texto, representado, por exemplo, por indicadores estruturais.

Esse módulo é baseado num POS tagger e na gramática de estado finito para identificar termos médicos e extrair definições. Na identificação de frases nominais simples foram utilizados um etiquetador e o analisador sintático parcial (chunker) de sintagmas nominais.

2) Módulo de análise gramatical que utiliza uma rica gramática lexicalista orientada à dependência para analisar fenômenos lingüísticos mais complexos, como aposição e anáfora.

4.2 Métodos Estatísticos

Os métodos baseados em conhecimento estatístico geralmente utilizam a freqüência de palavras para detectar candidatos a termos. Outro detalhe do método estatístico é a sua dependência no tamanho do corpus. Se o corpus é pequeno, poucos termos são encontrados do total existente, e mesmo quando o corpus é grande, alguns termos ficam de fora, pois têm baixa freqüência.

Além disso, sistemas estatísticos geram bastante ruído, que são candidatos a termos que não possuem valor terminológico, ou seja, são palavras pertencentes à língua geral.

Outra característica que diferencia o método estatístico do método lingüístico é a independência de língua.

Métodos puramente estatísticos para a extração automática de termos são pouco utilizados, pois a maioria dos métodos implementados são os híbridos. As principais medidas utilizadas no método estatístico de extração automática de termos são os seguintes:

- Freqüência de ocorrência (Daille, 1996; Manning & Schütze, 1999): É a medida mais simples para se encontrar termos em corpus. Se duas palavras aparecem freqüentemente juntas, existe a possibilidade de ela representar uma função especial. No entanto, existe o risco de palavras funcionais, como artigos e preposições, serem selecionadas como candidatos a termos. Este método é independente de domínio e recursos externos, mas ele é fraco para recuperar termos com baixa freqüência.

- Informação mútua (Pantel & Lin, 2001): É uma medida de informação que uma variável contém sobre outra, sendo ela a redução da incerteza de uma variável randômica devido ao conhecimento da outra. Os cálculos são feitos usando a probabilidade das palavras ocorrerem num corpus de tamanho N. Esta medida foi usada inicialmente para extração de *colocações*. Quando palavras ou termos são adjacentes, a informação mútua é maior. Da mesma forma que a freqüência de ocorrência, esta medida não é eficiente para extrair termos com baixa freqüência.
- Coeficiente *log-likelihood* (Pantel & Lin, 2001): Esta medida foi utilizada para amenizar o problema de recuperação de termos de baixa freqüência da medida mutua. A diferença dela para a informação mutua, é que ela também considera termos raramente adjacentes como fortes candidatos a termos.
- Coeficiente *Dice:* É uma medida similar à informação mútua, que faz um cálculo para saber a o nível de relação entre duas palavras. A diferença desta medida para a informação mútua, é que esta não depende do tamanho da amostra.
- C-value (Frantzy & Ananiadou, 1997): Esta medida é utilizada para extração de termos compostos. Ela é baseada na freqüência de cadeias no corpus, porém ela utiliza um cálculo para evitar que subcadeias de cadeias já extraídas sejam consideradas candidatas a termos.

4.3 Métodos Híbridos

Os sistemas híbridos combinam conhecimentos lingüísticos com conhecimentos estatísticos, e conseguem os melhores resultados. A ordem em que os conhecimentos são aplicados influencia os resultados. Sistemas em que o método estatístico é aplicado depois do lingüístico são os que geram melhores resultados, pois o conhecimento estatístico funciona como um complemento para o lingüístico.

4.4 O Trabalho de Teline (2004)

Teline (2004) fez uma avaliação de métodos das três abordagens de extração automática de termos, descritos nas seções prévias. Foi utilizado um corpus contendo 164 artigos da área de Revestimentos Cerâmicos. O corpus analisado possuía 448.352 palavras. A lista de referência de termos possuía 381 termos (264 unigramas, 74 bigrams e 43 trigramas).

Na abordagem estatística foram usadas as medidas de Freqüência para unigramas, Freqüência, Informação Mútua, Log-Likelihood e Coeficiente Dice para bigramas e para trigramas foram usadas Freqüência, Informação Mútua, Log-Likelihood. Em seguida foi feita uma análise manual da lista de unigramas e bigramas candidatos a termos. Esta intervenção foi feita com o intuito de eliminar palavras e siglas da língua geral, marcas publicitárias, nomes próprios e símbolos especiais. Neste caso, o método estatístico com intervenção humana é considerado um método semi-automático.

Na abordagem lingüística foi feito um pré-processamento no corpus para permitir a realização de consultas sobre o mesmo. O primeiro tipo de consulta realizada no corpus é a busca por expressões e indicadores estruturais, que são expressões lingüísticas que geralmente vem acompanhados de definições, descrições e outros tipos de orações que concentram termos. Em seguida é feita uma busca padrões morfossintáticos (*Ex. Substantivo + adjetivo*). Além disso, nesta abordagem é feito o uso de uma *stoplist*, que é uma lista contendo palavras da língua geral.

Na abordagem híbrida, foi feito o processamento do corpus e a buscas por expressões e indicadores estruturais, em seguida foram aplicados métodos estatísticos (cálculo de freqüência para unigramas, bigramas e trigramas e informação mútua para bigramas). Após essas etapas, foi feita uma intersecção com a lista de padrões morfossintáticos.

Para avaliar o desempenho dessas abordagens, foram utilizadas métricas clássicas da área de processamento de sinais, como a Precisão e Cobertura. **Precisão** é a razão das

respostas corretas recuperadas pelo sistema e todas as respostas recuperadas e **Cobertura** é a razão de respostas corretas e todas as respostas corretas possíveis.

Dentre os métodos implementados, o que retornou a melhor precisão foi o método semi-automático (estatístico + intervenção humana), seguido pela abordagem híbrida, estatística e por ultimo, a lingüística. Os resultados da precisão podem ser observados na Figura 5.

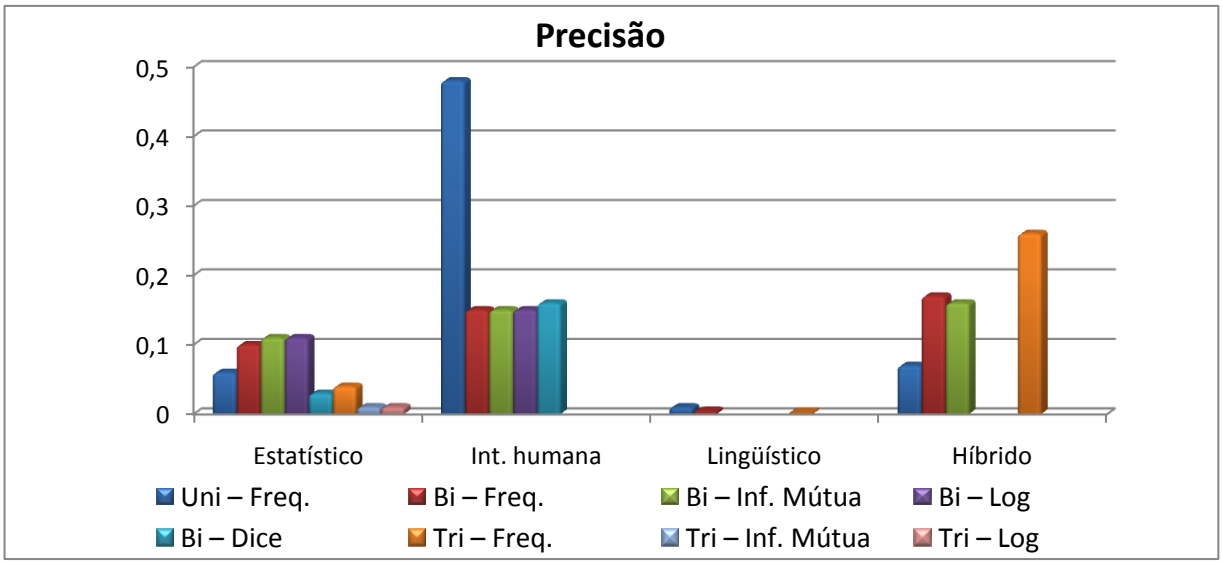

Figura 5 - Precisão obtida pelas três abordagens e a semi-automática (Teline, 2004)

Para cobertura, a abordagem lingüística retornou os melhores resultados, seguidos pela semi-automática, estatística e híbrida. Os resultados de cobertura obtidos podem ser observados na Figura 6.

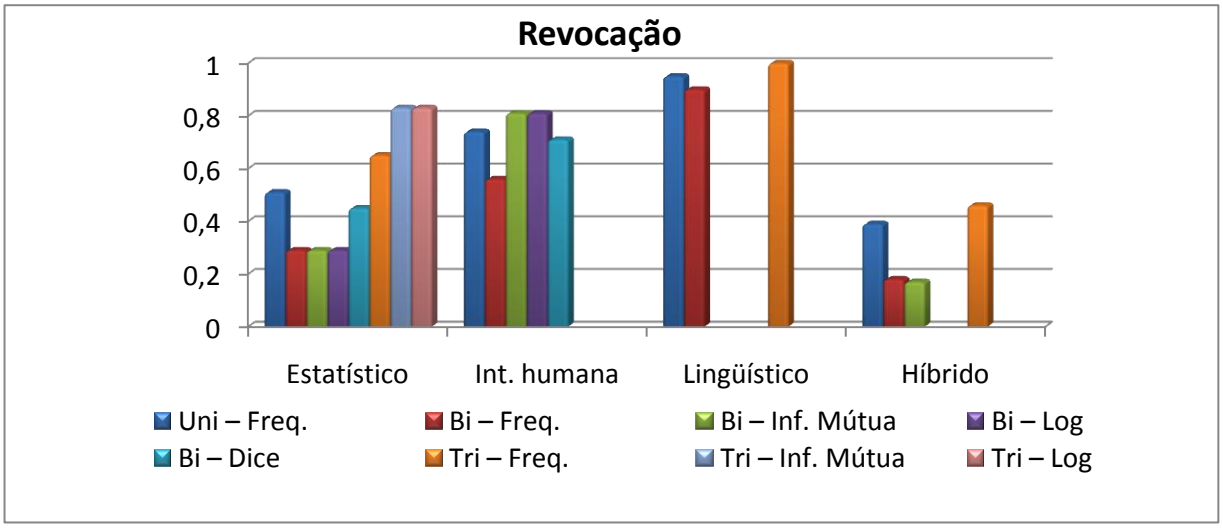

Figura 6 – Cobertura obtida pelas três abordagens e a semi-automática (Teline, 2004)

Para calcular o desempenho da tarefa, é usada a medida F, que é uma medida que permite colocar ênfase na precisão e na cobertura, sendo B um parâmetro que representa a importância relativa da precisão e cobertura:

$$F = \frac{(B^2 + 1) * precisão * revocação}{B^2 * (precisão + revocação)}$$

Novamente, o método semi-automático retornou os melhores resultados para a medida F, seguido da abordagem híbrida, estatística e lingüística, respectivamente. Os valores obtidos podem ser observados na figura 7.

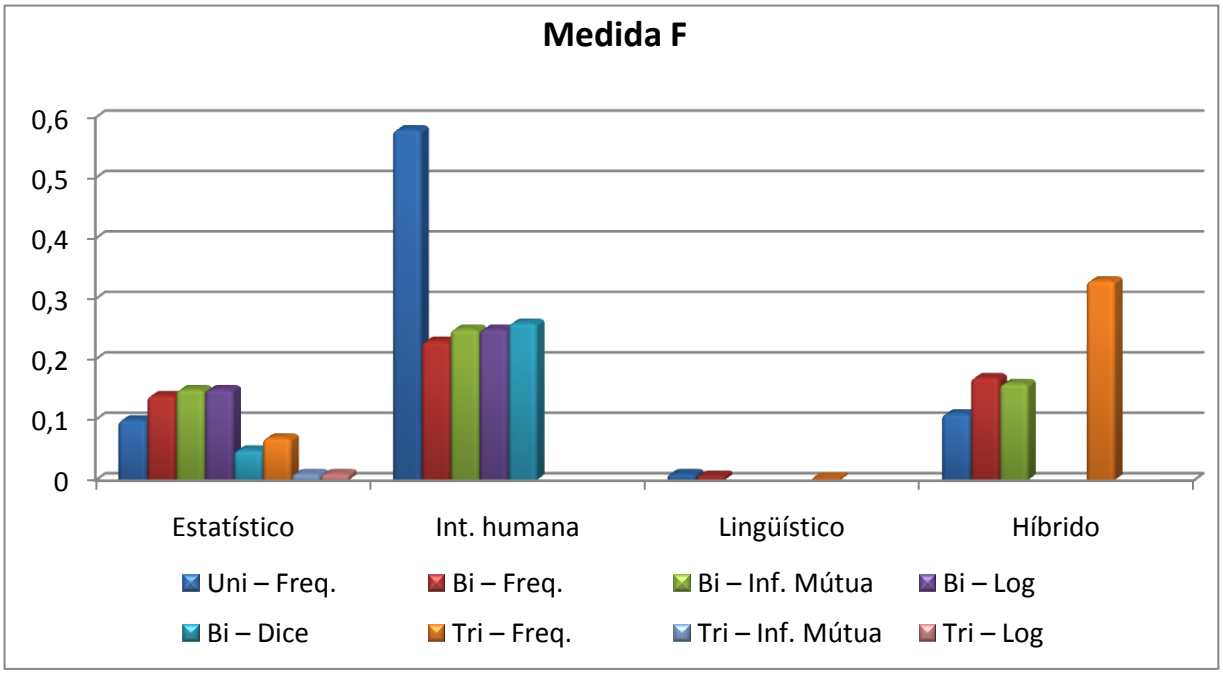

Figura 7 – Medida F obtida pelas 3 abordagens e a semi-automática (Teline, 2004)

A partir da análise dos resultados, concluiu-se que o método estatístico depende fortemente de uma intervenção humana para gerar resultados satisfatórios, e que o fato da medida F não ser tão alto (Abaixo da média de 60%, que é o normal para tarefa) se justifica pelo fato de que o corpus foi compilado a partir de uma única fonte.

4.5 O Trabalho de Ribeiro Jr (2008)

No trabalho sobre construção de ontologias de Ribeiro Jr (2008) foi necessário fazer uso da extração automática de termos. Neste caso, foi utilizada uma abordagem híbrida. Os conhecimentos lingüísticos utilizados foram a análise da classe gramatical, de sintagmas nominais e o uso de padrões morfossintáticos.

Em seguida, foram aplicados cálculos de relevância de freqüência, tf-idf (Medida que considera relevantes os termos que possuem alta freqüência de ocorrência em numero limitado de documentos) e NC-Value, que estão descritos em (Ribeiro Jr, 2008).

Para avaliar o método, foi utilizado um corpus da área de Ecologia (Zavaglia *et al*., 2007) constituído por textos extraídos de partes dos livros "A Economia da Natureza" e "Ecologia", além de revistas presentes no projeto LácioWeb[21]. O corpus conta com um total de 260.921 palavras.

Na extração de unigramas, houve um empate de desempenho entre o uso das classes gramaticais e o núcleo do sintagma nominal combinado com o cálculo NC-Value usando tf-idf como parâmetro de freqüência obtiveram os melhores índices de precisão (14,7%), cobertura (49,96%) e medida F (22,39%).

Para extração de bigramas, o melhor método foi o uso de padrões morfossintáticos com o cálculo de relevância de freqüência, obtendo índice de precisão de 5,7%, cobertura de 41,91% e medida F de 10,04%.

Na extração de trigramas, a melhor estratégia foi o uso de padrões morfossintáticos com o cálculo de relevância *tf-idf*, obtendo índice de precisão de 2,9%, cobertura de 46,77% e medida F de 5,46%.

Outra abordagem feita por (Ribeiro Jr, 2008) foi a adição do uso de conhecimento semântico extraído através do parser PALAVRAS (Bick, 2000) no método de extração híbrido. Nesta abordagem, é necessária a intervenção humana para escolher quais grupos semânticos serão utilizados na extração de termos, o que pode influenciar no resultado dependendo do nível de conhecimento do domínio da pessoa que fizer essas escolhas. Nos experimentos feito por (Ribeiro Jr, 2008), o uso de conhecimento semântico trouxe um aumento médio de 17,33% na precisão de extração de unigramas, aumento de 20,67% na precisão de extração de bigramas e um aumento de 6,66% na precisão de extração de trigramas.

[21] http://www.nilc.icmc.usp.br/lacioweb/

5 Projeto NorMan

O objetivo deste projeto de pesquisa chamado de **NorMan – Normalização de Manuais** é analisar como as relações entre ações *generation* (gera) e *enablement* (habilita) são realizadas em instruções de manuais técnicos do português e criar uma ferramenta de simplificação de manuais técnicos de produtos tecnológicos, baseada no editor de simplificação do PorSimples (SIMPLIFICA), adaptada ao gênero . Embora Delin *et al*. (1994) tenha realizado um estudo de corpus em português sobre estas relações (descrito na Seção 2.2.2.1.1) este estudo foi muito pequeno. Pretendemos estender o estudo e fazê-lo para 3 classes manuais diferentes: manuais de escopo internacional, nacional e local.

A seguir será apresentada a metodologia a ser utilizada para o desenvolvimento deste trabalho, bem como as tarefas realizadas e futuras.

5.1 Metodologia Baseada em Corpus para a Extração de Termos de Textos Instrucionais e para a Normalização (Simplificação) de Instruções

A execução deste trabalho será dividida em 4 etapas:

- Compilação e Análise de um corpus de manuais técnicos;
- Proposta de um Método de Extração de Termos técnicos sensível ao gênero instrucional/procedimental;
- Adaptação do Manual de Simplificação Sintática para tratar de textos instrucionais/procedimentais
- Adaptação do Editor de Simplificação SIMPLIFICA para manuais técnicos.

As etapas descritas acima serão detalhadas nas subseções seguintes.

5.1.1 Compilação e Análise de um Corpus de Manuais Técnicos

O primeiro passo a ser realizado neste projeto é a compilação de um corpus de manuais técnicos. A partir deste corpus poderá ser feito um estudo de como as relações entre ações *generation* (gera) e *enablement* (habilita) são realizadas.

O corpus a ser criado terá três conjuntos de manuais:

- Manuais técnicos de produtos fabricados por empresas de atuação mundial;
- Manuais técnicos de produtos fabricados por empresas de atuação nacional;
- Manuais técnicos de produtos fabricados por empresas locais.

A partir destes três conjuntos de manuais, esperamos obter uma amostra atualizada dos formatos de manuais utilizados por empresas de porte pequeno até as multinacionais.

Esta etapa do projeto já foi iniciada. Foi feita uma pesquisa na web e foi constatada a disponibilidade pública dos manuais de instrução da empresa Britânia[22], que é uma empresa brasileira com mais 50 anos de mercado e que comercializa eletro-portáteis e eletrônicos. Os manuais desta empresa se encaixam na categoria de manuais de empresa de atuação nacional.

Foram encontrados 333 manuais de instruções em formato pdf. Para tornar possível o trabalho com estes manuais, é necessário convertê-los para o formato de texto puro. Durante o processo de conversão, 144 manuais apresentaram problemas que ainda não foram resolvidos, restando 189 manuais em formato texto.

Além disso, o corpus será utilizado na tarefa de extração de termos, que será apresentada na Seção 5.1.2.

5.1.2 Proposta de um Método de Extração de Termos Técnicos Sensível ao Gênero Instrucional/Procedimental

Durante a tarefa de simplificação de manuais técnicos, é preciso tomar um cuidado especial em relação aos termos técnico presentes no texto. Manuais técnicos contêm termos que não podem ser excluídos. Caso um termo técnico seja erroneamente suprimido do texto ou trocado por um mais simples durante o processo de simplificação, o sentido final da sentença poderá ficar seriamente afetado, prejudicando o entendimento do leitor. Para evitarmos esse tipo de situação, será executada uma etapa de extração de termos, descrita nesta subseção.

Nesta etapa do projeto, o método de extração de termos implementado será uma **adaptação** do método híbrido implementado por Teline (2004). Neste método, o primeiro passo é a extração das orações que possuem certos padrões que são passados como parâmetro. Esses padrões são expressões e indicadores estruturais que geralmente vêm acompanhados de

[22] http://www.britaniashop.com.br/

definições, descrições, classificações e de outros tipos de orações que concentram termos. São utilizadas as expressões classificadas como de Uso Geral (UG), que são expressões que podem ser utlizadas em qualquer domínio e Sinais Gráficos (SG), que são indicadores como "()", ":" e "-" (Teline, 2004). Essas expressões foram retiradas de várias fontes (Aluísio,1995 apud Teline, 2004), (Sager, 1993 apud Teline,2004), (Klavans e Muresan 2000; 2001a; 2001b apud Teline, 2004) e também da lista ISO/TC 37/SC 1 Terminologia – Princípios e Métodos[23] e do CórpusEco, que é um corpus composto por textos do gênero científico, domínio da Ecologia, Botânica, Biologia Geral, Zootecnia, Recursos Florestais e Engenharia Florestal. Este corpus foi desenvolvido no NILC para o Projeto Bloc-Eco[24]. Essas expressões podem ser vistas na Figura 8.

Aluísio (1995)		**CórpusEco**	
apresenta	(UG)	adição de	
atua		chamamos	(UG)
caracterizado	(UG)	constitui	(UG)
classe de	(UG)	constituído	(UG)
compreendendo		depende	(UG)
compreendido	(UG)	desenvolvido	(UG)
conhecido como	(UG)	determinado	(UG)
consiste	(UG)	empregado	(UG)
contém, contêm	(UG)	expresso	(UG)
em outras palavras		formado	(UG)
implica	(UG)	obtido	(UG)
isto é	(UG)	palavra	(UG)
ou seja	(UG)	relacionado	(UG)
por exemplo	(UG)		
tal como	(UG)	**Klavans e Muresan (2000;2001a;2001b)**	
Sager (1993)		chamado	(UG)
é	(UG)	definido como	(UG)
são	(UG)	expressão	(UG)
utilizado	(UG)	(se) entende	
		significa	
ISO/TC 37/SC 1	(UG)	termo	(UG)
conceito	(UG)	( )	(SG)
corresponde		-	(SG)
define	(UG)	:	(SG)
denominado			
feito de	(UG)		
usado			

Figura 8 - Expressões linguísticas que serão utilizadas para a extração de termos

[23] http://www.tc37sc4.org/
[24] http://nilc.icmc.usp.br/nilc/projects/bloc-eco.htm

O próximo passo é a aplicação de medidas estatísticas utilizando o pacote NSP (N-gram Statistics Package)[25], escrito em Perl. O pacote NSP foi implementado por Ted Pedersen, Satanjeev Banerjee e Amruta Purandare na Universidade de Minnesota, Duluth. Ele é constituído por um conjunto de programas que auxilia na análise de n-gramas em arquivos texto.

Após o cálculo das medidas estatísticas, os termos extraídos são filtrados por uma lista de padrões morfossintáticos, gerando a lista final de termos. Os padrões morfossintáticos para extração de termos compostos são os mesmos usados por Baségio (Baségio, 2006 apud Ribeiro Jr, 2008), apresentados na Figura 9.

Substantivo Adjetivo Preposição Adjetivo Substantivo Adjetivo
Substantivo Adjetivo Preposição Adjetivo Substantivo
Substantivo Preposição Adjetivo Substantivo Adjetivo
Substantivo Preposição Adjetivo Substantivo
Substantivo Adjetivo Preposição Substantivo Adjetivo
Substantivo Adjetivo Preposição Substantivo
Substantivo Preposição Substantivo Adjetivo
Substantivo Preposição Substantivo
Substantivo Adjetivo

Figura 9 - Padrões morfossintáticos para extração de termos compostos

Para termos simples, são utilizados os padrões usados por (Teline, 2004), que podem ser observados na Figura 10.

Substantivo
Adjetivo
Verbos

Figura 10 - Padrões morfossintáticos usados na extração de termos simples

Na adaptação do método híbrido, os padrões da Figura 9 serão trocados pelos padrões das relações gera e habilita, extraídos da análise de corpus realizada na tarefa descrita na Seção 5.1.1. Os padrões morfossintáticos serão, muito possivelmente os mesmos mostrados nas Figuras 9 e 10.

[25] http://www.d.umn.edu/~tpederse/nsp.html

5.1.3 Adaptação do Manual de Simplificação Sintática e do SIMPLIFICA para Tratarem de Textos Instrucionais/Procedimentais

Nesta etapa do projeto, com a análise de corpus realizada, serão propostas regras de simplificação para as relações gera e habilita e será implementada uma interface para edição de manuais técnicos baseada no Editor de Simplificação do projeto PorSimples.

A partir da interface, o usuário poderá criar textos para manuais técnicos, e durante essa etapa de criação do texto, serão apresentadas sugestões e correções visando à simplificação do texto instrucional.

Para isso, será feito uso do parser PALAVRAS (Bick, 2000) para a extração do conhecimento morfossintático dos textos, e através desse conhecimento, serão identificados os casos de simplificação apresentados no manual de simplificação, já adaptado. Após a identificação da regra, o sistema aplica as regras específicas para cada caso e então a sugestão de simplificação com a sua devida explicação será retornada para a interface do usuário e assim o mesmo pode avaliar se aplica ou não a regra.

5.2 Tarefas Realizadas e Tarefas Futuras

Abaixo são listadas as tarefas já realizadas e as que devem ser realizadas até março de 2009, para a finalização do mestrado:

1. Durante o primeiro e segundo semestre deste projeto, foram cursadas as disciplinas:

Teoria da Computação; Tipos e Estruturas de Dados; Introdução a Inteligência Artificial; Sistemas Inteligentes; Tópicos em Processamento de Língua Natural; Seminário em Computação I e II; Preparação Pedagógica.

2. Além disso, também já foi realizado o exame de proficiência em línguas (Inglês).

3. Realização de pesquisa bibliográfica relacionada ao tema de trabalho.

4. Preparação e defesa da monografia de qualificação.

5. Realizar a compilação e análise do corpus de textos instrucionais.

6. Adaptar o Manual de Simplificação.

7. Projeto e Implementação do Editor de Simplificação do projeto NorMan.

8. Analisar, testar e avaliar o Editor com usuários reais.

9. Escrita de artigos, relatórios técnicos, dissertação e defesa.

5.3 Cronograma

A Tabela 3 apresenta as nove atividades propostas para o projeto.

Tabela 3 - Cronograma de atividades

	2009										2010		
Tarefa	**Jan-Mar**	**Abr**	**Mai**	**Jun**	**Jul**	**Ago**	**Set**	**Out**	**Nov**	**Dez**	**Jan**	**Fev**	**Mar**
3	X	X	X	X	X	X	X	X	X	X			
4	X												
5	X	X	X	X									
6				X	X								
7					X	X	X	X					
8									X	X			
9									X	X	X	X	X

Referências

ALTWARG, R. (2006). Controlled Languages, an Introduction. Centre for Language Technology website on Controlled Natural Languages, Macquarie University. Available at http://www.shlrc.mq.edu.au/masters/students/raltwarg/clindex.htm

ALUÍSIO, S.; SPECIA, L.; PARDO, T.; MAZIERO, E.; CASELI, H. M.; FORTES, R. (2008a). "A Corpus Analysis of Simple Account Texts and the Proposal of Simplification Strategies: First Steps towards Text Simplification Systems " In the proceedings of The 26th ACM Symposium on Design of Communication (SIGDOC 2008), pp. 15-22.

ALUÍSIO, S.; SPECIA, L.; PARDO, T.; MAZIERO, E.; FORTES, R. (2008b). "Towards Brazilian Portuguese Automatic Text Simplification Systems. " In the proceedings of The Eight ACM Symposium on Document Engineering (DocEng 2008), pp. 240-248.

AOULADOMAR, F. (2005a). Towards Answering Procedural Questions, Workshop KRAQ05, IJCAI05, Edinburgh, 2005.

AOULADOMAR, F. (2005b). A Preliminary Analysis of the Discursive and Rhetorical Structure of Procedural Texts, Symposium on the Exploration and Modelling of Meaning, SEM-05, Biarritz (Basque Country, France), 2005.

AOULADOMAR, F., SAINT-DIZIER, P. (2005). An Exploration of the Diversity of Natural Argumentation in Instructional Texts, Workshop CMNA, IJCAI05, Edinburgh, July 2005.

BICK, E. (2000). The Parsing System “Palavras”. Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. Tese (Doutorado) – Arhus University, 2000.

BIDERMAN, M. T. C. (2005). DICIONÁRIO Ilustrado de Português. São Paulo, Editora Ática.. 1ª. ed. São Paulo: Ática, 2005.

BURG J. F. M., Linguistic Instruments in Requirements Engineering, IOS Press, Amsterdam, The Netherlands, 1996

CARVALHO NETTO, J. R. (2003). Ao Encontro da Lei: O Novo Código Civil ao alcance de todos. São Paulo: Imprensa Oficial, 2003.

COLE, R. (1997). Survey of the state of the art in human language technology, Cambridge University Press, New York, NY, 1997

DAILLE, B. (1996). Study and Implementation of Combined Techniques for Automatic Extraction of Technology. In: Klavans, J., Resnik, P. The Balancing ACT- Combining Symbolic and Statistical Approaches to Language, The MIT Press, p. 49-66.

DELIN, J.; HARTLEY, A.; PARIS, C., SCOTT, D. ; VANDER LINDEN, K. (1994). Expressing Procedural Relationships in Multilingual Instructions, *Proceedings of the Seventh International Workshop on Natural Language Generation*, pp. 61-70, Maine, USA. 1994.

Di EUGENIO, B. (1992). Understanding Natural Language Instructions: the Case of Purpose Clauses, Proceedings 30th Annual Meeting of the Association for Computational Linguistics, ACL92, Newark, Delaware, 120-127, 1992. Disponível em http://www.aclweb.org/anthology-new/P/P92/P92-1016.pdf

EDWARDS, J. A. and A. G. KINGSCOTT. (1997). eds. *Language Industries Atlas*. Second edn. (Amsterdam: IOS Press, 1997).

EIJK, P.(1997), Controlled languages in technical documentation (1997) Computational Linguistics in the Netherlands, 1997 – pag 187

ESTOPÀ BAGOT, R. (1999). Extracció de terminologia: elements per a la construcció d'um SEACUSE (Sistema d'Extracció Automàtica de Candidats a Unitats de Signficació Especialitzada). Tese de Doutorado. Universidade Pompeu Fabra.

FELLBAUM, C. (1998). *WordNet An Electronic Lexical Database*, The MIT Press, 1998.

FONTAN, L., SAINT-DIZIER, P. (2008). *Analyzing the explanation structure of procedural texts: dealing with Advices and Warnings.* Dans / In : *International Symposium on Text Semantics (STEP 2008), Venise, 22/09/2008-24/09/2008*, Johan Bos (Eds.), Association for Computational Linguistics (ACL), p. 84-93, septembre / september 2008.

FRANTZY, K. T.; ANANIADOU, S. (1997). Automatic Term Recognition using Contextual Cues. Manchester Metropolitan University. THIRD DELOS WORKSHOP Cross-Language Information Retrieval Zurich, 5-7 March 1997 ISBN 2-912335-02-7.

GRISHMAN, R. (2001). Adaptive Information Extraction and Sublanguage Analysis. Proceedings of IJCAI,2001. Disponível em: http://nlp.cs.nyu.edu/publication/papers/grishman-ijcai01.pdf

HEID, U.; JAUβ, S.; KRÜGER, K; HOHMANN, A (1996). Term extraction with standard tools for corpus exploration. IN: 4th International Congresso n Terminology and Knowledge Engineering. Wien. August.

JANCZURA, G.A. ; CASTILHO, G. M. ; ROCHA, N. O. ; VAN ERVEN, T.J.C.; HUANG, T. P. (2007). Normas de concretude para 909 palavras da língua portuguesa. Psicologia. Teoria e Pesquisa, v. 23, p. 195-204, 2007.

KAJI, H. (1999). "Controlled languages for machine translation: State of the art," Proceedings of Machine Translation Summit VII, pp.37-39, 1999.

KLAVANS, J. L.; MURESAN, S. (2000). DEFINDER: Rule-Based Methods for the Extraction of Medical Terminology and their Associated Definitions from Online Text. In: Proceedings of AMIA 2000.

KLAVANS, J. L.; MURESAN, S. (2001a). Evaluation of DEFINDER: A System to Mine Definitions from Consumer-oriented Medical Text. In: Proceedings of JCDL 2001.

KLAVANS, J. L.; MURESAN, S. (2001b). Evaluation of the DEFINDER System for Fully Automatic Glossary Construction. In Proceedings of AMIA 2001.

LEFFA, V.J. (1996). Fatores da compreensão na leitura. Cadernos no IL, Porto Alegre, v.15, p.143-159, 1996. <http:// www.leffa.pro.br/fatores.htm)>. Acesso em jan. 2008.

MANNING, C.; SCHÜTZE, H. (1999). Collocations. In: Foundations of Statistical Natural Language Processing, p. 141-77. MIT Press. Cambrige.

MAPLESON, D.L. (2006). Post-Grammatical Processing for Discourse Segmentation. PhD Thesis. School of Computing Sciences, University of East Anglia, Norwich.

MAX, A. (2006). Writing for Language-impaired Readers. InProceedings of Seventh International Conference on Intelligent Text Processing and Computational Linguistics (Mexico City, Mexico, February 19-25, 2006). CICLing 2006. Springer-Verlag, Berlin Heidelberg New York, 567-570.

MITKOV, R. (2003). The Oxford Handbook of Computational Linguistics (Oxford Handbooks in Linguistics S.), Oxford University Press, 2003

OGDEN, C.K. (1932). *Basic English, A General Introduction with Rules and Grammar*. London: Paul Treber & Co.

PANTEL, P.; LIN, D. (2001). A statistical corpus-based term extractor. In: E. Stroulia & S. Matwin (Ed.), AI 2001, Lecture Notes in Artificial Intelligence, Springer-Verlag, p. 36–46.

PARIS, C. and SCOTT, D. (1994). Stylistic variation in multilingual instructions. In Proceedings of the Seventh International Workshop on Natural Language Generation, Kennebunkport, MN, 21--24 June 1994, pages 45--52.

PARIS, C.; VANDER LINDEN, K., FISCHER, M.; HARTLEY, A.; PEMBERTON, L.; POWER, R.; SCOTT, D. (1995). A support tool for writing multilingual instructions. In Proceedings of the 14th International Joint Conference on Artificial Intelligence, pages 1398--1404, Montreal, Canada, 1995.

PETERSEN, S. E.; OSTENDORF, M. (2007). Text Simplification for Language Learners: A Corpus Analysis. 2007. In Proceedings of the Speech and Language Technology for Education Workshop (Pennsylvania, USA, October 1-3, 2007). SLaTE-2007. Carnegie

Mellon University and ISCA Archive, http://www.isca-speech.org/archive/slate_2007. 69-72.

RIBEIRO JR., L. C. (2008). OntoLP : Construção Semi-Automática de Ontologias a partir de Textos da Língua Portuguesa. Dissertação (Mestrado) - Programa de Pós-Graduação em Computação Aplicada, Universidade do Vale do Rio dos Sinos, 2008.

RIBEIRO, V. M. (2006). Analfabetismo e alfabetismo funcional no Brasil. Boletim INAF. São Paulo: Insitituto Paulo Montenegro

SAINT-DIZIER, P. (1998). *Verb Semantic Classes Based on 'Alternations' and WordNet-like criteria*, in : Predicative Forms in Natural language and lexical Knowledge Bases, Reds: Saint-Dizier,P., Eds: Kluwer Academic, Cambridge, USA, 1998.

SIDDHARTHAN, A. (2002). An Architecture for a Text Simplification System. In the Proceedings of the Language Engineering Conference (LEC), pp. 64-71. (2002)

SPECIA, L.; ALUISIO, S.M.; PARDO, T.A.S. (2008). "Manual de Simplificação Sintática para o Português". Technical Report NILCTR0806, 27 p. Junho 2008, São Carlos -SP.

TELINE, M. F. (2004). Avaliação de métodos para extração automática de terminologia de textos em português. ICMC-USP, São Carlos, 2004. Dissertação de Mestrado.

WILLIAMS, S. (2004). Natural Language Generation (NLG) of discourse relations for different reading levels. PhD Thesis, University of Aberdeen. (2004)

ZAVAGLIA, C. *et al*. (2007). Estrutura ontológica e unidades lexicais: uma aplicação computacional no domínio a ecologia. In: Anais do 5º Workshop em Tecnologia da Informação e da Linguagem Humana, TIL'2007. Rio de Janeiro, Brasil: [s.n.], 2007. P. 1575-1584.

Apêndice A

Lista de 850 palavras do vocabulário do Inlês Básico. Elas denotam conceitos simples e são usadas para situações do dia a dia.

Operações - 100 palavras

come, get, give, go, keep, let, make, put, seem, take, be, do, have, say, see, send, may, will, about, across, after, against, among, at, before, between, by, down, from, in, off, on, over, through, to, under, up, with, as, for, of, till, than, a, the, all, any, every, no, other, some, such, that, this, I, he, you, who, and, because, but, or, if, though, while, how, when, where, why, again, ever, far, forward, here, near, now, out, still, then, there, together, well, almost, enough, even, little, much, not, only, quite, so, very, tomorrow, yesterday, north, south, east, west, please, yes.

Coisas - 400 palavras gerais

A-F

account, act, addition, adjustment, advertisement, agreement, air, amount, amusement, animal, answer, apparatus, approval, argument, art, attack, attempt, attention, attraction, authority, back, balance, base, behaviour, belief, birth, bit, bite, blood, blow, body, brass, bread, breath, brother, building, burn, burst, business, butter, canvas, care, cause, chalk, chance, change, cloth, coal, colour, comfort, committee, company, comparison, competition, condition, connection, control, cook, copper, copy, cork, cotton, cough, country, cover, crack, credit, crime, crush, cry, current, curve, damage, danger, daughter, day, death, debt, decision, degree, design, desire, destruction, detail, development, digestion, direction, discovery, discussion, disease, disgust, distance, distribution, division, doubt, drink, driving, dust, earth, edge, education, effect, end, error, event, example, exchange, existence, expansion, experience, expert, fact, fall, family, father, fear, feeling, fiction, field, fight, fire, flame, flight, flower, fold, food, force, form, friend, front, fruit

G-O

glass, gold, government, grain, grass, grip, group, growth, guide, harbour, harmony, hate, hearing, heat, help, history, hole, hope, hour, humour, ice, idea, impulse, increase, industry, ink, insect, instrument, insurance, interest, invention, iron, jelly, join, journey, judge, jump, kick, kiss, knowledge, land, language, laugh, law, lead, learning, leather, letter, level, lift, light, limit, linen, liquid, list, look, loss, love, machine, man, manager, mark, market, mass, meal, measure, meat, meeting, memory, metal, middle, milk, mind, mine, minute, mist, money, month, morning, mother, motion, mountain, move, music, name, nation, need, news, night, noise, note, number, observation, offer, oil, operation, opinion, order, organization, ornament, owner

P-Z

page, pain, paint, paper, part, paste, payment, peace, person, place, plant, play, pleasure, point, poison, polish, porter, position, powder, power, price, print, process, produce, profit, property, prose, protest, pull, punishment, purpose, push, quality, question, rain, range, rate, ray,

reaction, reading, reason, record, regret, relation, religion, representative, request, respect, rest, reward, rhythm, rice, river, road, roll, room, rub, rule, run, salt, sand, scale, science, sea, seat, secretary, selection, self, sense, servant, sex, shade, shake, shame, shock, side, sign, silk, silver, sister, size, sky, sleep, slip, slope, smash, smell, smile, smoke, sneeze, snow, soap, society, son, song, sort, sound, soup, space, stage, start, statement, steam, steel, step, stitch, stone, stop, story, stretch, structure, substance, sugar, suggestion, summer, support, surprise, swim, system, talk, taste, tax, teaching, tendency, test, theory, thing, thought, thunder, time, tin, top, touch, trade, transport, trick, trouble, turn, twist, unit, use, value, verse, vessel, view, voice, walk, war, wash, waste, water, wave, wax, way, weather, week, weight, wind, wine, winter, woman, wood, wool, word, work, wound, writing, year.

Coisas - 200 palavras concretas

angle, ant, apple, arch, arm, baby, bag, ball, band, basin, basket, bath, bed, bee, bell, berry, bird, blade, board, boat, bone, book, boot, bottle, box, boy, brain, brake, branch, brick, bridge, brush, bucket, bulb, button, cake, camera, card, cart, carriage, cat, chain, cheese, chest, chin, church, circle, clock, cloud, coat, collar, comb, cord, cow, cup, curtain, cushion, dog, door, drain, drawer, dress, drop, ear, egg, engine, eye, face, farm, feather, finger, fish, flag, floor, fly, foot, fork, fowl, frame, garden, girl, glove, goat, gun, hair, hammer, hand, hat, head, heart, hook, horn, horse, hospital, house, island, jewel, kettle, key, knee, knife, knot, leaf, leg, library, line, lip, lock, map, match, monkey, moon, mouth, muscle, nail, neck, needle, nerve, net, nose, nut, office, orange, oven, parcel, pen, pencil, picture, pig, pin, pipe, plane, plate, plough, pocket, pot, potato, prison, pump, rail, rat, receipt, ring, rod, roof, root, sail, school, scissors, screw, seed, sheep, shelf, ship, shirt, shoe, skin, skirt, snake, sock, spade, sponge, spoon, spring, square, stamp, star, station, stem, stick, stocking, stomach, store, street, sun, table, tail, thread, throat, thumb, ticket, toe, tongue, tooth, town, train, tray, tree, trousers, umbrella, wall, watch, wheel, whip, whistle, window, wing, wire, worm.

Qualidades - 100 palavras descritivas

able, acid, angry, automatic, beautiful, black, boiling, bright, broken, brown, cheap, chemical, chief, clean, clear, common, complex, conscious, cut, deep, dependent, early, elastic, electric, equal, fat, fertile, first, fixed, flat, free, frequent, full, general, good, great, grey, hanging, happy, hard, healthy, high, hollow, important, kind, like, living, long, male, married, material, medical, military, natural, necessary, new, normal, open, parallel, past, physical, political, poor, possible, present, private, probable, quick, quiet, ready, red, regular, responsible, right, round, same, second, separate, serious, sharp, smooth, sticky, stiff, straight, strong, sudden, sweet, tall, thick, tight, tired, true, violent, waiting, warm, wet, wide, wise, yellow, young.

Qualidades - 50 opostos

awake, bad, bent, bitter, blue, certain, cold, complete, cruel, dark, dead, dear, delicate, different, dirty, dry, false, feeble, female, foolish, future, green, ill, last, late, left, loose, loud, low, mixed, narrow, old, opposite, public, rough, sad, safe, secret, short, shut, simple, slow, small, soft, solid, special, strange, thin, white, wrong.